17568

NATIONAL LIBRARY
OTTAWA

BIBLIOTHÈQUE NATIONALE
OTTAWA

CANADA

NAME OF AUTHOR.......Mr Wanlop Kansup

TITLE OF THESIS.......A Comparison of Several Methods of Assessing
Partial Knowledge in Multiple Choice Tests

UNIVERSITY.......The University of Alberta

DEGREE FOR WHICH THESIS WAS PRESENTED.......M Ed

YEAR THIS DEGREE GRANTED.......1973

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed).......

PERMANENT ADDRESS:

64 M.15, Soi Sukin 3, Petkasem
Highway, Bangduan, Parsicharoen,
Bangkok 6, Thailand

DATED.......27 August.......1973

NL-91 (10-68)

THE UNIVERSITY OF ALBERTA

A COMPARISON OF SEVERAL METHODS OF ASSESSING PARTIAL KNOWLEDGE IN MULTIPLE CHOICE-TESTS

by

© WANLOP KANSUP

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN
PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF EDUCATION
IN
TESTING AND MEASUREMENT

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA
FALL, 1973

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled "A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests " submitted by Wanlop Kansup in partial fulfilment of the requirements for the degree of Master of Education in Testing and Measurement.

....

Supervisor

V. R. Nyberg

H Kass

Date August 10, 1973

ABSTRACT

The use of multiple-choice test items involving one choice out of four or five alternatives has been so extensive that many test writers and users now employ no other form of test. There are, however, two major disadvantages inherent in the conventional one-or-zero scoring system. The first of these is that this method is unable to discriminate between partial information, complete information, and no information. The second undesirable feature is the encouragement of guessing.

In the present study an attempt was made to improve upon current techniques of administering and scoring the multiple-choice test, or, more specifically, to increase the reliability and validity of such tests over what it is under conventional administration and scoring.

Three test-taking methods were used in this study: Conventional Testing, Confidence Testing, and Elimination Testing. Four different scoring techniques were used along with these test-taking methods. Conventional Scoring was used with Conventional Testing and Elimination Testing; Differential Weighted Scoring, with Conventional Testing; Confidence Weighted Scoring, with Confidence Testing; and Elimination Scoring, with Elimination Testing. The Confidence Weighted Scoring was done by use of five different scoring functions, three of which were introduced in this

study.

There were two aptitude tests--vocabulary and mathematics--used with these experimental methods. Two types of criterion scores were obtained: school achievement and aptitude test scores from similar forms of the vocabulary and mathematics aptitude tests. The aptitude tests as criteria were administered and scored using conventional procedures.

Subjects were 1028 grade nine students randomly assigned to three groups of comparable size. Each group was given the tests under one of three test-taking methods. For each group, there were two test sessions. In the first session two tests were given--vocabulary and mathematics. In the second, the same two tests were given again, and two other aptitude tests were administered, using the conventional treatment, to obtain test scores for use as criteria. School achievement scores were obtained from school records.

It was found that two scoring functions employed with the Confidence test-taking method provided scores more reliable than did the conventional method with either Conventional or Differential Weighted Scoring. These two functions are based on the increasing increment scoring model. Both the functions and the scoring model were introduced in this study. It was found, however, that none of the experimental scoring techniques provided test scores more valid than provided by the conventional test-taking and scoring methods.

Since validity is generally the most important test characteristic, the findings indicate that the experimental test-taking and scoring approaches may not be worth the effort. Discussion of the findings, in relation to the results of some previous studies and to theoretical implications, are given. It is suggested that, in order to solve the shortcomings inherent in the use of the conventional test-taking and scoring procedures, investigators might be wise to pursue other leads than those examined in the present study.

ACKNOWLEDGEMENTS

The writer wishes to thank the members of the thesis committee, Dr. A. R. Hakstian, Dr. V. R. Nyberg and Dr. H. J. Kass, for their contribution to the completion of this study. Special thanks are expressed to Dr. S. M. Hunka for his untiring assistance in data processing.

Grateful acknowledgement is extended to the school boards, [illegible] principals, teachers, and all grade IX and university students involved in the project.

Particular appreciation is extended to his friends: J. M. Richmond, for his valuable advice and his assistance in data collection; R. G. Baril, for his many constructive suggestions and help in data processing, and E. Skakun, for his help in APL programming.

Gratitude is expressed to the Canadian International Development Agency for its financial support, without which this project could not have been undertaken.

Finally, the writer wishes to thank his wife, Siwalai, and his children, Wasina and Parima, for their forebearance and encouragement during the two years of his absence from home.

TABLE OF CONTENTS

Chapter Page

I. BACKGROUND OF THE STUDY 1
  Introduction to the Problem 1
  Statement of the Problem and Implications for Education. 5
  Limitation of the Study 6
  Research Hypotheses 7
  Theoretical Framework 8
    Definition of Terms. 8
    Basic Assumptions. 11

II. RELATED LITERATURE. 13
  The Differential Weighting of Item Responses. 14
  The Response-Determined Scoring 18
  Personality Influences on Test Scores. 24
  Theoretical Facet of the New Techniques. 26
  Conclusion. 30

III. RESEARCH METHODOLOGY. 31
  Design of the Study 31
  Subjects. 32
  Instruments 33
  Test-Taking Instructions. 37

Chapter Page

Test Administration Procedures 38

A Technique for Differential Weighting of Item Options. 39

Scoring Techniques 41

Problems With the Confidence Scoring Functions. 44

A Pilot Analysis on the Value of log 0 44

Possible Item Scores from Different Scoring Functions. 47

School Achievement 55

Data Obtained to Test Hypotheses 55

Summary of Testing Design and Data Sets. 57

Statistical Hypotheses 57

Statistical Techniques 58

IV. RESULTS 61

Reliability 63

Coefficient of Internal Consistency 63

Summary of Results. 71

Test-Retest Reliability 71

Summary of Results. 78

Validity. 78

School Achievement Scores as Criteria. 79

VB and RB Test Scores as Criteria. 96

Chapter Page

Summary of Results 115

V. CONCLUSIONS. 116

Summary. 116

Findings and Implications. 117

Suggestions for Further Research 127

BIBLIOGRAPHY 129

APPENDIX A: TEST INSTRUCTIONS, ANSWER SHEETS, AND OPTION WEIGHTS. 134

APPENDIX B: NORMS FOR VOCABULARY TEST: FORM B AND MATHEMATICS APTITUDE TEST: FORM B. 154

APPENDIX C: INTERCORRELATIONS, MEANS AND STANDARD DEVIATIONS. 158

LIST OF TABLES

TABLE Description Page

1. NUMBER OF STUDENTS IN EACH GROUP ACCORDING TO SCHOOLS 34

2. NORMALIZED STANDARD SCORES OF r_{ij} ON THE DISTRIBUTION OF 11 POSSIBLE VALUES OF r_{ij}. . . 43

3. THREE SETS OF POSSIBLE ITEM SCORES FROM DIFFERENT VALUES OF LOG 0. 46

4. RELIABILITIES (ALPHA COEFFICIENTS) OF TEST SCORES FROM DIFFERENT VALUES OF LOG 0. . . . 46

5. POSSIBLE ITEM SCORES FROM FIVE SCORING FUNCTIONS. 48

6. DIFFERENCES BETWEEN TWO SUCCESSIVE POSSIBLE ITEM SCORES FROM FIVE SCORING FUNCTIONS. . . 48

7. NUMBERS ASSIGNED TO LETTER-GRADES FOR STANDARDIZING. 56

8. TESTING DESIGN FOR VA AND RA TESTS AT BOTH TESTING TIMES 58

9. A SUMMARY OF DATA SETS IN THE STUDY. 59

10. ALPHA COEFFICIENTS IN CV GROUP 64

11. ALPHA COEFFICIENTS IN CF GROUP 64

12. RANK ORDERS OF ALPHA VALUES FROM THE SAME TEST. 66

13. ALPHA COEFFICIENTS IN EL GROUP 66

14. CRITICAL VALUES OF THE F-RATIO WITH DIFFERENT DEGREES OF FREEDOM FOR TESTS OF DIFFERENCES BETWEEN ALPHA COEFFICIENTS ACROSS GROUPS. 68

15. F-RATIO OF COMPARISONS OF ALPHA COEFFICIENTS FROM VA_1 BETWEEN SCORING TECHNIQUES ACROSS GROUPS 69

TABLE Description Page

16. *F*-RATIO OF COMPARISONS OF ALPHA COEFFICIENTS FROM RA$_1$ BETWEEN SCORING TECHNIQUES ACROSS GROUPS 69

17. TEST-RETEST RELIABILITIES IN CV GROUP. 72

18. TEST-RETEST RELIABILITIES IN CF GROUP. 73

19. RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF VA TEST SCORES. 73

20. RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF RA TEST SCORES. 74

21. RANK ORDERS OF RELIABILITIES WITHIN EACH TEST. 74

22. TEST-RETEST RELIABILITIES IN EL GROUP. 75

23. RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF VA TEST SCORES ACROSS GROUPS. 76

24. RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF RA TEST SCORES ACROSS GROUPS 77

25. INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN CV GROUP 80

26. INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN CF GROUP 80

27. INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN EL GROUP 80

28. VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN CV GROUP. 82

29. DIFFERENCES BETWEEN VALIDITIES UNDER DWS AND CVS$_1$ SCORING TECHNIQUES. 82

30. VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN CF GROUP. 84

31. DIFFERENCES BETWEEN VALIDITIES OF VA$_1$ TEST SCORES WITH LA. 85

32. DIFFERENCES BETWEEN VALIDITIES OF VA$_2$ TEST SCORES WITH LA. 85

TABLE Description Page

33. DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH MA 86

34. DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH MA 86

35. DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH SC 87

36. DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH SC 87

37. DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH AV 88

38. DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH AV 88

39. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH LA 89

40. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH LA 89

41. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH MA 90

42. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH MA 90

43. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH SC 91

44. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH SC 91

45. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH AV 92

46. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH AV 92

47. VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN EL GROUP 95

48. DIFFERENCES BETWEEN VALIDITIES UNDER ELS AND CVS_2 SCORING TECHNIQUES 95

49. CORRELATIONS OF VB AND RB TEST SCORES WITH SCHOOL ACHIEVEMENT. 96

TABLE Description Page

50. VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN CV GROUP 98

51. DIFFERENCES BETWEEN VALIDITIES UNDER DWS AND CVS_1 SCORING TECHNIQUES. 98

52. VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN CF GROUP 99

53. DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH VB. 100

54. DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH VB. 100

55. DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH RB. 101

56. DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH RB. 101

57. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH VB. 102

58. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH VB. 102

59. DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH RB. 103

60. DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH RB. 103

61. VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN EL GROUP 105

62. DIFFERENCES BETWEEN VALIDITIES UNDER ELS AND CVS_2 SCORING TECHNIQUES. 105

63. VALIDITIES OF VA_1 AND RA_1 TEST SCORES UNDER SEVEN SCORING TECHNIQUES 107

64. RANK ORDERS OF THE VALIDITIES WITH THE SAME CRITERION. 109

65. COMPARISONS OF THE VALIDITIES OF VA_1 TEST SCORES WHEN SCHOOL ACHIEVEMENT SCORES WERE CRITERIA. 109

TABLE	Description	Page
66.	COMPARISONS OF THE VALIDITIES OF RA_1 TEST SCORES WHEN SCHOOL ACHIEVEMENT SCORES WERE CRITERIA	110
67.	COMPARISONS OF THE VALIDITIES OF VA_1 TEST SCORES WHEN VB AND RB WERE CRITERIA	111
68.	COMPARISONS OF THE VALIDITIES OF RA_1 TEST SCORES WHEN VB AND RB WERE CRITERIA . .	112
69.	RESULTS FROM SOME PREVIOUS STUDIES COMPARING EXPERIMENTAL METHODS WITH CONVENTIONAL METHOD.	120
70.	COMPARISONS OF THE VALIDITIES OF VA TEST SCORES UNDER THE CONVENTIONAL AND CONFIDENCE METHODS	126
71.	COMPARISONS OF THE VALIDITIES OF RA TEST SCORES UNDER THE CONVENTIONAL AND CONFIDENCE METHODS	126
72.	NUMBER OF UNIVERSITY STUDENTS RATING FOR OPTION WEIGHTS	151
73.	OPTION WEIGHTS FOR VOCABULARY TEST: FORM A	152
74.	OPTION WEIGHTS FOR MATHEMATICS APTITUDE TEST: FORM A	153
75.	NORMS FOR VOCABULARY TEST: FORM B.	155
76.	NORMS FOR MATHEMATICS APTITUDE TEST: FORM B	157
77.	INTERCORRELATIONS AMONG SCORES IN CONVENTIONAL TESTING GROUP	159
78.	INTERCORRELATIONS AMONG SCORES IN CONFIDENCE TESTING GROUP	160
79.	INTERCORRELATIONS AMONG SCORES IN ELIMINATION TESTING GROUP.	161
80.	MEANS AND STANDARD DEVIATIONS OF TEST SCORES UNDER CONVENTIONAL TESTING AND SCORING METHODS.	162

LIST OF FIGURES

Figures Page

1. A Plot of x_m against m 49

2. A plot of d_n against n 50

CHAPTER I

BACKGROUND OF THE STUDY

Introduction to the Problem

The use of multiple-choice test items involving one choice out of four or five proffered alternatives has been so extensive that many test writers now use no other form (Hendrickson, 1971). But the general acceptance of the multiple-choice form of test item as the best one for objective measurement of aptitude or achievement does not imply that it has reached its optimal form. Any variation upon an already widely accepted technique which indicates promise of improved measurement is deserving of further investigation (Coombs, Milholland, & Womer, 1956).

During the last two decades or more there has been an increasing interest in looking for new methods of scoring this form of test in place of the conventional one-or-zero method. There have been a number of suggestions for new complex scoring patterns for objective tests, particularly multiple-choice ones (Thorndike & Hagen, 1969, p. 123). All these ideas stemmed from the fact that test writers and users, at present, are still unsatisfied with the conventional scoring method commonly employed. Among the several disadvantages of the conventional procedure, the most seemingly crucial points are: (1) the inability to discriminate between partial information and complete information or misinformation, and (2) the encouragement of guessing (Coombs et al., 1956). Accordingly, investigators have spent much time and effort finding out whether there are other reliable techniques to replace the old one, hoping that the new techniques might provide the possibility of eliminating these disadvantages and result in increased test reliability and validity. Scoring formulas that assess partial knowledge of examinees have been proposed and used in many studies. The conclusion of these studies is that partial knowledge does exist and that by employing proper scoring techniques the reliability of multiple-choice tests may be increased (Coombs, et al., 1956; Sabers & White, 1969).

Among several alternatives in studies to assess partial knowledge, two kinds of techniques are typically employed: those that (1) differentially weight the response alternatives, and (2) require examinees to report their confidence in the correctness of response alternatives (Hambleton, Roberts, & Traub, 1970). For the first method, some studies carried out in comparison with the conventional method have reported that there is a certain amount of increment in test reliability and also test validity (Davis & Fifer, 1959; Dressel & Schmid, 1953; Hambleton et al., 1970; Rippey, 1970; Sabers & White, 1969). However, because the increments they found were sometimes small and sometimes inconsistent, there are still no sound conclusions that can be drawn from those works with respect to the applicability of the new technique. Several investigators are still optimistic about the problem and have suggested that further studies be employed (Hambleton et al., 1970; Sabers & White, 1969; Stanley & Wang, 1970).

More studies involving the use of confi[illegible]ce weighting have been reported than those using the differential weighting technique. Two reasons which seem to contribute to the evidence are the promising results of the technique found in earlier works and the variety of formulas suggested for assigning the confidence level. Among those studies reported, most show an increase of test reliability and some an increase in test validity, although many of them could not find any consistency in the increment (Dressel & Schmid, 1953; Ebel, 1965; Hopkins, Hakstian, & Hopkins, 1973; Michael, 1968). At present there are still no definitive conclusions about the contribution of the new techniques to testing. However, although these investigators could not formulate definitive conclusions from their works, they did recommend ways of improving future studies (Hambleton et al., 1970).

Another technique suggested in the literature which is supposed to assess partial knowledge is the method of elimination (Coombs et al., 1956). Although this technique is considered as one of the confidence weighting methods (Wang & Stanley, 1970), the procedure seems rather distinctive. In this method subjects are instructed to eliminate incorrect alternatives, taking care not to mark the correct one. Item score depends on (1) option weights, +1 for each incorrect alternative crossed out and -3 for the correct alternative crossed out in a four-choice item, and (2) the examinee's confidence level as shown by the number of alternatives he crossed out. There is one study by Coombs _et al_.(1956) that used this technique and its results indicated a potentially promising contribution to the testing field.

Because of the dissatisfaction with the disadvantages of the conventional scoring and the promising results from the new techniques reported in various studies, investigators have tried to approach the problem with different and improved designs to obtain more precise results. It was generally hoped that if these studies proved the superiority of the new techniques over the old one there would be a radical change in testing practices. None of the many studies reported in the literature, however, has demonstrated definitely the advantages of a new technique over the old one. The problem still exists and awaits solution.

It should be noted that there is at least one study that comes very close to providing some sort of definitive conclusions. This work was done by Hambleton _et al_. (1970). They used both differential weighting and confidence weighting techniques in comparison with the conventional technique in the same study. The main purpose was to see whether significant increments in both test reliability and validity could be found. Because of the fact that the number of subjects was small and the test was rather easy for the subjects' level, they failed to establish the expected conclusion. However, the failure did not discourage them from showing enthusiastic hope and optimism about the new techniques.

The most recent work reported in the literature on this problem was done by Hopkins et al. (1973). This study was concentrated on the confidence weighting. The investigators tried to improve on the procedures previously employed. They were successful in showing an increment in test reliability with confidence weighting but failed to find any consistent evidence for increased test validity. This work should not be regarded as providing closure to the problem, because of the different design and method of scoring they used in their study.

It can be concluded that all studies reported thus far, although worth the effort and contributing to the solution of the problem, are still not sufficient to make final definitive conclusions. Further more comprehensive studies are required before sound conclusions can be made.

With the ideas discussed above and an optimistic view about the new techniques, the author decided to solve the problem with a more comprehensive experimental design.

Statement of the Problem and Implications for Education

As many studies have reported, the basic objective of the investigation is to discover a new method to score multiple-choice test items that can eliminate two crucial disadvantages resulting from the use of the conventional method: (1) the inability to assess partial knowledge, and (2) the encouragement of guessing. Theoretically, it is widely agreed that these new techniques can eliminate both disadvantages. But even though they can, it does not imply that these techniques are ready to be used. Improvements in test reliability and validity must be demonstrated, and empirical evidence is required in this regard.

In this study, effort was made to assess whether improvement in reliability and validity resulted from the use of different test-taking and scoring methods. More procedures were assessed, and a larger sample was used than in earlier work. It was hoped that, by utilizing a more comprehensive design, more definitive results would be obtained.

The above discussion leads to the main implication for educational practice, that if this study showed significant improvements in test reliability and validity for any of the techniques compared, the results would be significant to the multitudes of test users. With the already accepted ability of these techniques to assess partial information and eliminate guessing, and with demonstrated improved reliability and validity of test scores, the methods should be found attractive by test users.

Limitation of the Study

This study was confined to grade nine junior high school students. Therefore, the results are representative of subjects at this school level only, and should be cautiously generalized to higher or lower levels of education, i.e., elementary school, senior high school, or college.

The tests used in this study were aptitude tests; verbal comprehension and arithmetic reasoning, and the criteria were school achievement scores at the end of the first term. So inference should, strictly speaking, be confined to these predictors and criteria.

No attempt was made to compare experimental and conventional methods with respect to administration time, levels of test difficulty, levels of intelligence, sexes, or any other variables not previously discussed.

Research Hypotheses

The ultimate goal of this study was to examine differences in test reliability and validity using the new testing and scoring methods, using as a baseline results obtained by administering and scoring tests under conventional procedures. It was expected that test scores obtained under either experimental method would contain more information about the state of each examinee's knowledge than the conventional one since they allow partial information to be taken into account.

From the discussed rationale, this study was carried out to examine the following hypotheses:

(1) Test scores obtained under any of the experimental methods are more reliable than those obtained under the

conventional method. Evidence examined would include both the internal consistency of test scores and the stability of test scores over a period of time.

(2) Test scores obtained under any of the experimental methods have higher validity for outside criteria measuring either (a) the same or similar traits or (b) school achievement, than those obtained under the conventional method.

Theoretical framework

(1) Definition of Terms

(a) Testing Methods.

Conventional Testing (CV Method). This is the test-taking procedure most widely used at present. The procedure is to encourage examinees to read a question and then choose the one of four or five response alternatives that they judge to be the correct answer.

In this study subjects were encouraged to answer all items. However, no explicit instructions to guess were given.

Confidence Testing (CF Method). The technique was a modified version of that used by Michael (1968), and Hambleton *et al*. (1970). In this technique, each subject was instructed to distribute 10 points of confidence among five response alternatives for an item according to the correctness he thought each one had.

Elimination Testing (EL Method). This test-taking method was developed and used by Coombs *et al*. (1956). In this method a subject is asked to select and mark incorrect answers out of five given response alternatives for an item. He can mark one or two up to four given answers leaving one that he thinks is correct. The item score depends on the number of choices he marks and whether he marks the correct answer or not.

(b) Scoring Methods.

Conventional Scoring (CVS Method). This scoring method was applied to answers under both Conventional Testing and Elimination Testing. Under the Conventional Testing, a correct answer would score one, an incorrect answer, including omissions, zero. Under the Elimination Testing format, a score of one was given to the answer with all four incorrect alternatives crossed out, otherwise, including omissions, zero was given. Thus an item score would be either one or zero. The total test score was the sum of the item scores. No correction for guessing was applied.

Differential Weighted Scoring (DWS Method). This scoring method was used by Hambleton et al. (1970). In this study it was applied to answers under the Conventional Testing format. The technique was different from the Conventional Scoring in that, instead of scoring one or zero, Differential Weighted Scoring assigned predetermined weights to all incorrect response alternatives according to the degree of correctness each one had, with a weight of five to the correct answer. The item score was the weight of an option chosen by a subject and the total test score was the sum of the item scores.

In this study, the weight of each response alternative was assigned on the basis of a pilot analysis using a group of students in the Department of Educational Psychology at the University of Alberta. Details of this pilot study are given in Chapter III.

Confidence Weighted Scoring (CWS Method). This scoring method was used with true-false test items by Ebel (1965) and then discussed in details by Shuford, Albert, & Massengill (1966). It was then used by other investigators with some modifications (Hambleton et al., 1970; Hopkins et al., 1973; Michael, 1968). In this scoring method, either the original point a subject gives to the correct answer is taken as the item score or a function (usually non-linear) of this value is used. In this study the technique was applied to answers under the Confidence Testing format. Five scoring functions, including two suggested and used in previous studies, were used to get different item scores. The total test score was the sum of item scores obtained by each scoring function. Details of the scoring functions are given in Chapter III.

Elimination Scoring (ELS Method). This scoring method was used by Coombs et al. (1956) and was applied to answers under the Elimination Testing format in this study. The technique was slightly modified for use with a five-choice item. Each incorrect answer selected scored one;

the correct answer, if selected, scored -4; omissions scored 0. An item score was an algebraic sum of all response alternatives marked ranging from -4 to +4. The total test score was the algebraic sum of all the item scores.

(c) School Achievement. This term was used with grades or scores obtained from schools at the end of the first term, of the 1972-73 school year. These scores were used as criteria in this study. Grades were obtained for the following subjects:

(a) Language Arts.

(b) Mathematics.

(c) Science.

The academic average was obtained from scores of all subjects listed above plus Social Science.

Basic Assumptions. For the purpose of objective interpretations that would be drawn from this study, it was assumed that:

(1) All test and criterion scores used in this study were reliable with respect to Conventional Testing and Scoring.

(2) Subjects used in this study were only randomly different from one another across all test format groups.

(3) Subjects clearly understood test instructions given at the testing times.

(4) Subjects were cooperative and eager to obtain high test scores.

(5) There were no differences between the experimental

and conventional test format groups with respect to the following factors: administration time, subjects' and proctors' personalities, intelligence, and sex.

CHAPTER II

RELATED LITERATURE

When measures are to be combined to form a composite measure or to predict a criterion, the question of differential item or subtest weighting arises. There is actually a theoretical rationale to support this idea. The need for differential item-option weighting generally arises from the desire to improve the reliability and validity of a composite scores. The term differential weighting has been used for several techniques. Some investigators have weighted tests, some test items, and others item responses. Some even went beyond this by developing the method of response-determined scoring which can also be described as a form of differential weighting (Wang & Stanley, 1970).

Although differential weighting of item or item-option theoretically promises to provide substantial gain in test reliability and validity, in practice, some approaches employed often gain so slightly that they do not seem to justify the labour involved in deriving the weights and then employing them in scoring. A number of psychologists have concluded that some of these weighting approaches, especially the item-weighting type, are not worth the trouble (Guilford, 1954; Gulliksen, 1950). However, the differential item-option weighting on aptitude or achievement tests has shown potential, and it has been proposed by test specialists that reliability and validity of a test may be increased if a subject himself assigns weights to options according to his confidence in the correctness of each option (Wang & Stanley, 1970).

There are several distinct techniques of *option* weighting reported in the literature and some of them, with modification, have been used in this study. Each will be reviewed separately and then treated as an integrated whole.

The Differential Weighting of Item Responses

Differential weighting techniques differ from the conventional technique in that, instead of scoring one for a correct answer and zero for an incorrect or omitted response, weights (usually *a priori*) are assigned to each response alternative to an item. A test score then consists of the sum of the weights of the response alternatives that the subject chose in responding to the test (Hambleton *et al*., 1970). A wide variety of techniques has been proposed and used to assign weights to response alternatives. Until fairly recently, the possibility of the differential weighting of item response was not considered in the literature (Wang & Stanley, 1970). However, after the demise of formula scoring which tried to eliminate the effect of guessing on test scores, some investigators have turned their attention to this new approach. The first step in the direction of differential weighting of incorrect response was made by Nedelsky (1954). He used the opinions of experts to identify distractors with respect to the achievement levels of subjects. In spite of the complicated technique he used, the composite score was considerably more reliable than the conventional score.

Davis & Fifer (1959) took a significant second step after Nedelsky. These authors noted that the conventional one-or-zero scoring did not permit differentiation among examinees with respect to the type of distractors selected. Students should also be differentiated by the option they select. This idea really concerns the assessment of partial knowledge. Two subsequent steps were made. The first emphasized test reliability and the second, test validity. The tests used in this study were two forms of arithmetic reasoning consisting of 50 items each. Option weights were obtained empirically by three successive steps. First, two mathematicians rated all options with respect to their degrees of correctness on a seven-point scale. The weights were then used to score subjects' answers in a pilot study. The second step was to find a new set of weights using the correlation between marking the option and the total score obtained by the first set. Option weights from this step were then modified and adjusted to form the final set for the main study. This complex procedure produced a satisfactory change in reliability but no increase in validity. The test-retest reliability of test scores under the experimental scoring method was .763 as compared with .684 obtained under the conventional method, and the difference between the two values attained statistical significance. But there was no significant increase in the validity coefficient. The authors' conclusion is that the variance introduced into the total score had increased the proportion of true variance, thus increasing test reliability, but that the new variance displayed the same concurrent validity as the original, thus resulting in the unchanged test validity.

Jacob & Vandeventer (1968) undertook a study using the notion of facet analysis (Guttman & Schlesinger, 1967) to obtain option weights on the *Coloured Progressive Matrices Test* (CPM). This procedure made use of the mean number of respondents choosing the distractor as an option weight. This *a priori* method of keying the response alternatives was shown to have a moderate degree of test-retest reliability, and concurrent and predictive validity (Hambleton *et al*., 1970; Wang & Stanley, 1970).

Sabers & White (1969) used differential weighting with 370 grade nine students divided into four groups. The experimental test was the *Iowa Algebra Aptitude Test* (IAAT) which was given to students while in the eighth grade. The upper and lower 27 per cent of a group were chosen on the basis of scores on an achievement test. The percentage within these groups marking an item option on the IAAT were used to obtain the weight for that option from the table prepared by Davis (1966). Weights obtained by this procedure from one group were used to score tests for another group in order to assess the cross-validity of the weighted scoring.

The criterion measures were 40-item multiple-choice achievement tests administered after the students had completed one semester in ninth-grade mathematics. The achievement tests were scored as the number correct. Sabers & White found that differential weighting resulted in small increments in both the reliability and predictive validity. They postulated that two factors contributed to the failure to achieve larger increments. These were: (1) the groups were not well matched on their aptitudes, and (2) the aptitude test had a relatively high degree of reliability.

Hambleton *et al*. (1970) compared two *a priori* methods of option weights, one based on the procedure of Jacob & Vandeventer (1968) and one based on the average rank assigned to options by judges who ranked all options for correctness. The experimental method was used to score a five-option midterm test and the criterion score was a final examination consisting of a 60-item multiple-choice test administered and scored under the conventional method. This study also included the confidence method which will be discussed later. The results showed that both sets of weights tended to increase estimated predictive validity of the midterm examination, and the reliability was slightly increased with the second set of weights. None of these increments of the validity attained statistical significance and no test of the differences between reliabilities was made. However, the authors noted that, in this study, the number of subjects in each group was rather small and that the test used was easy for the group level being tested.

The most recent study on differential weighting was done by Hendrickson (1971). She used option weights secured by using Guttman's technique for maximizing internal consistency. The weights were derived via an iterative procedure which began by assigning to a response alternative a weight equal to the mean total score on the remaining items of the sub-test obtained by the examinees who marked that response alternative. The technique was applied to a large number of subjects who took the Scholastic Aptitude Test (SAT), Verbal and Mathematics forms, totalling 100,000 persons. Hendrickson found that weighting options resulted in increased reliability equal to lenghtening the test from 19.09 per cent to 78.25 per cent of the original item number under the conventional scoring, but there were no clear effects on test validity which was slightly decreased. Hendrickson concluded that weighted options resulted in increased homogeneity of the test thus changing what it is measuring, and thereby decreasing test validity.

The Response-Determined Scoring

All the weighting techniques discussed above have the characteristic of a constant multiplicative weight being directly associated with the response alternatives. Once the weights have been determined, the examinee's score on a test is completely determined by the response options he selects. Response-determined scoring represented by the elimination and confidence weighted scoring methods is an alternative strategy for obtaining item scores. The distinctive characteristics of this scoring approach is that the examinee's response to an item consists of more than simply selecting the correct option. Under this approach, the concept of a 'response' has been considerably broadened, and it is the characteristic of the response, rather than those of the item or the option, which determines the item score (Wang & Stanley, 1970).

There are a large number of studies reported in the literature using one form or another of this scoring approach. Some of them have used the elimination method and others, the confidence weighting method, as used in this study.

Dressel & Schmid (1953) studied four different experimental methods, two of which can be categorized as two of the methods under consideration in the present study: the Free-Choice Test and the Degree-of-Certainty Test. The Free-Choice test required students to mark as many choices as needed in order to be sure that they had not omitted the correct answer. Each incorrect mark was scored -1/4 point and the correct mark was scored one. The Degree-of-Certainty test, considered as one form of the confidence weighting rather than the elimination method, required students to indicate the degree of certainty they had in the single answer they selected by assigning one of four possible values. Item score for this latter method depended on whether they selected the incorrect answer or not, and on the degree of certainty they assigned, resulting in the range from -4 to 4. The test used in this study was a five-option test and subjects were college students. This combination of response method and system yielded reliabilities of .67 for the Free-Choice test and .73 for the Degree-of-Certainty test as compared with .70 obtained with the conventional scoring procedure, but no difference attained statistical significance. The conclusion from the results of these two experimental methods, as referred to by Echternacht (1972), was that . . .

> . . . superior students, defined in term of traditional test scores, differed significantly from average and poor students when using the free-choice format, the difference being that high performers marked fewer answers across each of three different levels of item difficulty. . . . The degree-of-certainty method, on the other hand, differentiated superior, average, and low-ability students about equally well, the confidence marks being about the same for both average and difficult items. It was also concluded that the certainty factor measured by the free-choice item was not the same as that measured by the degree-of-certainty item (p. 221).

Coombs et al. (1956), performed an experiment complementary to the Free-Choice test in which subjects were instructed to eliminate incorrect alternatives in four-choice test items, taking care not to mark the correct one. The technique is termed the Elimination Method in the present study. One point is gained for each incorrect alternative crossed out and three points lost if the correct response is crossed out resulting in item score ranging from -3 to 3. Three 40-item multiple-choice tests were used.

They were: a vocabulary test, a test of drive information, and a test of spatial visualization. Subjects were 855 junior and senior high school students, grouped under three methods of testing: the conventional method (C), the experimental method (E), and the 'both' method (B). These three groups were matched on aptitude as measured by several standardized tests. The main assumption of the study was that partial information exists and enters into answering multiple-choice items. The authors found that the reliability of tests under the experimental method showed increases equivalent to that produced by a 20 per cent increase in the length of a conventional test of the same type. The authors also pointed out that as the difficulty of the test increased, the reliability also increased, and that the same item discriminated well when administered in either multiple-choice or experimental formats.

Another method of assessing partial knowledge entails the use of confidence weighting or the personal probabilistic approach, as preferred by some authors (de Finetti, 1965; Rippey, 1968; Shuford et al., 1966). The general technique of this approach is to ask students to assign weights to all response alternatives indicating preference or degree of belief for each one. These assigned weights are then subjected to some pre-determined scoring functions to obtain item scores. This procedure has its historical antecedents in connection with the true-false format, in studies by Henver (1932) and Soderquist (1936). More recently, Ebel

(1965) found that confidence weighting could improve true-false test reliability.

Dressel & Schmid (1953) studied the confidence weighting technique with multiple-choice test items. With the weighting technique they used, a reliability of .73 was found using this experimental method as opposed to .70 for the conventional one. Michael (1968) studied the use of a 10-point-confidence distribution scheme with 432 senior high school students in history classes using the STEP:Social Studies test, comparing with conventional scoring and formula scoring (correction-for-guessing). She reported evidence of increased reliability from .764 to .840 when using the confidence testing. No test of differences between the two values was made. Her conclusion was that the confidence weighting method affords considerable promise in affecting a higher estimate of reliability and a lower standard error of measurement than does either the conventional or the formula method, and that the confidence weighting method was a workable technique that could be employed by the average classroom teacher.

Hopkins et al. (1973), used only three levels of confidence distribution, H, M, and L, in their study with 63 graduate students taking a statistic course. The test used was a 65-item multiple-choice test with a short-answer test measuring the same content as a criterion for validity. The results of this study showed an increased reliability from .883, by the conventional method, to .915, by the experimental method, but decreased validity from .701 to .661. However, these differences did not attain statistical significance. The findings led to the conclusion that . . .

> . . . the added reliable variance often observed in confidence testing studies may be irrelevant response style variance and does not increase validity, in fact, it may actually diminish validity (Hopkins et al., 1973, p. 140).

Studies discussed thus far have shown that differential weighting of distractors including the several forms of confidence weighting in aptitude and achievement tests, have been examined with a great interest. Rarely have studies used both the elimination and confidence weighting approaches in a single investigation. However, there are two pieces of work in which this has been done: one is by Hambleton et al. (1970) and the other by Collet (1971). In the study by Hambleton et al., the authors used both differential weighting and confidence weighting in the same study. The procedure for differential weighting has already been described. In the confidence weighting part, students were asked to distribute 100 points of degree of certainty among five response alternatives on a specially designed answer sheet. These subjects' confidence weights were then used to obtain item scores via a logarithmic function, the version which corresponds to the first logarithmic function used in the present study. The authors reported insignificant improvement in validity (.72 as compared with .62) for confidence weighting over conventional scoring, but decreased reliability (.655 as compared with .711). No test of differences between reliabilities was made.

Collet (1971) compared new scoring approaches with the correction-for-guessing method. This study used two experimental scoring methods, the differential weighting and the elimination which is one type of confidence weighting. The elimination method was the same as that used by Coombs _et al_. (1956), and the differential weighting method was similar to that used by Davis & Fifer (1959). Two 50-item multiple-choice tests were obtained from two parallel forms of the _Henman-Nelson Test of Mental Maturity_ (college level) and were given to six, 47-student groups of undergraduate students. Criterion scores for validity were obtained from the _Washington Pre-College Test_ administered some 18 months before the study. The results indicated that both reliability and validity obtained by the elimination method were higher than those obtained by the correction-for-guessing method. However, the results were reversed with the differential weighting method. Only the validity obtained by the elimination method was significantly different from the correction-for-guessing method. The author's conclusion was thus in favour of the elimination approach.

Personality Influences on Test Scores

Personality traits have long been suggested as possible factors influencing test scores under new scoring approaches (Coombs _et al_., 1956; Michael, 1968). However, there are only two studies reported in the literature that deal directly with these influences. Hansen (1971) studied the influence of variables other than knowledge on confidence or the probabilistic test scores, as he called them. Personality factors, in this study, were Rist Taking, Test Anxiety, and some others as measured by the F-scale developed by Christie, Havel, & Seidenberg (1958). The author found that the response style is related to certain aspects of personality.

Echternacht, Boldt, & Sellman (1972) attempted to evaluate the association of personality variables with confidence testing in light of practice. The testing technique that they used were called the "Distribute 100 points" and the "Pick-One" techniques. The first one required students to respond with subjective probabilities to each of the item alternatives, and the latter required students to select the best alternative and then rate their confidence in that choice on a five-point scale. Item scores were then computed using a logarithmic function similar to that suggested by Shuford et al. (1966). Subjects were 192 males in the U.S. Air Force. A personality test battery was developed from several well-known personality tests. The results showed some significant correlations between test scores and personality factors but the evidence did not hold up with replications. All of the findings led the authors to the conclusion that . . .

> . . . the personality variables are not related to confidence test scores when achievement, as measured by the number of items correctly answered, is controlled,

> and when sufficient practice with the system has been employed . . . grave skepticism about the use of confidence measures due to undue effects is probably not justified. At least, such reservations are not warranted any more for confidence measures than they are for traditional multiple choice measured (Echternacht et al., 1972).

Theoretical Facet of the New Techniques

In contrast to the many empirical studies on the assessment of partial knowledge, there are just a few dealing with theoretical aspects of these approaches. De Finetti (1965) introduced the use of Decision Theory into the problem of testing, calling it the Personal-Probability Approach. He suggested six preliminary assumptions which constitute the underlying philosophy of the approach. They were:

1. The scoring method and permitted modes of responding must be known to the subjects, the subjects fully understanding the implications in the face of uncertainty.

2. The subjects must be keenly interested in scoring high.

3. The subjects must be trained to understand the correspondence between their own belief and the numerical probabilities to which they were translated.

4. The total knowledge and belief of a given subject about a question and its alternatives must be expressed and fully represented by numerical probabilities he attached to each of the alternatives.

5. The scores using any scoring method can be divided so as to determine the partial information of a subject from his responses.

6. The evaluation of this procedure should concern how well the scoring method describes the subject's belief and its value to him (Echternacht et al., 1972).

In addition to these assumptions, de Finetti attempted to provide some rationale for behavior as he presented and discussed various scoring schemes.

Stokes (1966), in his article suggesting a new testing technique called "Split-Response Technique," viewed teachers' advantages as incentives for the use of a new approach. He suggested no theoretical implications of the technique. In his view, the Split-Response Technique will give more information to teachers especially in the following ways:

- Ambiguous otherwise poor questions can be detected by the frequency with which they give rise to split responses.
- Particular alternatives which are split testify to students uncertainty and point to meaningful test review items.
- Split responses on non-ambiguous questions serves as a barometer for ineffective teaching.
- Student confidence can be estimated better by observing the degree of splitting relative to the class.
- New freedoms in test design are possible, the teacher using questions with several correct alternatives which require splitting.
- New insight into student personality are possible (Stokes, 1966).

Shuford et al. (1966) gave a well known discussion on "Admissible Probability Measurement Procedures." Their objective was to extract a larger portion of the available information from objective test items. This information, as they stated, was contained in the student's degree-of-belief probabilities or personal probabilities concerning the correctness of the various possible answers. To measure these probabilities, they contended, a scoring system must be devised so that any student could maximize his expected score if, and only if, he honestly reported his probabilities. Scoring systems that made use of this property and were understood by students were termed admissible probability measurement procedures. In their view, most commonly used measurement procedures were not admissible. The authors also introduced the concept of a scoring system with a reproducing property; a scoring system was reproducible when the personal probabilities possessed by the examinee were identical to the probabilities with which he responded. They derived some necessary and sufficient conditions for the reproducibility of a test item with two possible alternatives. They further showed the class of reproducible scoring systems to be virtually inexhaustible and demonstrated a method of construction. However, all scoring functions they suggested and proved reproducible, except a logarithmic function, depended on both probabilities assigned to correct and incorrect alternatives. And because of the unbounded property of the logarithmic function when the probability assigned to the correct answer was zero, they suggested an approximation solution, a truncated logarithmic function, in which the value of -1 was given to log 0. This logarithmic function was used in several studies (Echternacht _et al_., 1972; Hambleton _et al_., 1970; Rippey, 1968, 1970).

Lord & Novick (1968) devoted an entire chapter to the problem of measurement procedures and item scoring formulas. They began by stating that the general problem of obtaining the maximum amount of information from a given set of items contains three major components. The first is the measurement procedure, or the manner in which the examinees respond to the item. The second is the specification of the item scoring rule or formula that is used for scoring each item. The third is the combination of item scores into a total score by an item weighting formula. The first two procedures are concerned directly with the problem of item scoring and the third one examines the problem of item weighting which is not being considered here. When dealing with the problem of choosing either simple or more complex measurement procedures, they suggested that . . .

> . . . it may be that examinees are available for a relatively long period for testing, but that test items are very difficult to obtain. Then we would want to obtain as much information as possible from each item, and hence we would be tempted to employ more complicated measurement procedures, if they were indeed useful. . . . It may be that items are plentiful but examinee time is scarce. It may also be reasonable to assume that per unit of time, we can probably get more information by adding more items (if available) than by introducing complex measurement procedures. Then we would probably be inclined to use the simpler measurement procedure so that we might administer as many items as possible in the limited amount of time (Lord & Novick, 1968, p. 303).

After a review of some possible scoring approaches, they advised that . . .

> . . . what little experimental work has been done on the traditional methods of formula scoring has not been encouraging, and that no experimental work has been published that supports the new methods. Thus, at present, the sole recommendation of these new methods is their strong conceptual attractiveness. In evaluating any new response method, it will be necessary to show that it adds more relevant ability variation to the system

> than error variation, and that any such relative increase in information retrieved is worth the effort, . . . (Lord & Novick, 1968, p. 314).

However, the authors preferred method seems to be that of the Personal-Probability Technique. They recommended that . . .

> That assumptions of the personal probability model are certainly more realistic than the assumption of the random guessing model. Many of the questions raised . . . may well be answered satisfactorily by empirical studies (Lord & Novick, 1968, p. 320).

Conclusion

This chapter has reviewed the literature on the problem of assessing partial knowledge, both empirical and theoretical. The literature indicates that partial knowledge can be measured and described some promising potential approaches. It also shows that further studies are needed. It is hoped that the present study will make, at least, a partial contribution to the clarification of the problems under consideration.

CHAPTER III

RESEARCH METHODOLOGY

Design of the Study

The present study employed three randomly assigned groups: Conventional Testing, Confidence Testing, and Elimination Testing. Subjects were 1028 grade nine students, of both sexes, selected from eight schools in Edmonton, Alberta and vicinity, during the first term of the 1972-73 school year. The testing instruments were two forms: A and B, of vocabulary and mathematics aptitude tests.

There were two testing times; the first was in October and the second in November and December 1972, with approximately three to five weeks between sessions for each class. At both testing times, students in three groups from the same classroom sat to write tests in the same room. They were given two form-A tests, vocabulary and mathematics aptitude tests, at two successive times, with one of three different test-taking instructions: Conventional Testing, Confidence Testing, and Elimination Testing, depending on which group they were assigned to.

In addition to form-A tests, form-B tests were given to students at the second test session after the first form in the same sequence: vocabulary and then mathematics aptitude, using Conventional Testing for all students.

Students' answers to form-A tests, using several scoring methods, provided data for test reliability. Answers to form-B tests, using the Conventional Scoring method, provided data for test validity.

At the end of the first term (three to five weeks after the second test session), school final scores from three subject areas: Language Arts, Mathematics, and Science, together with an Academic Average, were obtained from school files. These scores, after standardization, were used for test validity with respect to students' achievement in schools.

Details regarding subjects, tests, test administration procedures, scoring techniques, and school achievement, are presented in subsequent sections.

Subjects

Early in September 1972, five school systems in Edmonton and vicinity were requested to take part in the present study. Eight schools were willing to participate in the project. Two test sessions were then taken during the months of October and December 1972, with a 3-5 week interval between sessions for each class. The total sample, after excluding those whose scores were not complete, consisted of 1028 students.

Before the first test was taken, students in each class were randomly assigned into three comparably-numbered groups: the Conventional Testing, the Confidence Testing, and the Elimination Testing groups. The group division was made in each class for the purpose of eliminating class biases. However, during testing, all students sat at their usual places regardless of the group to which they belonged. The only difference among the three groups was that they were given different test-taking instructions.

No effort was made to equate these groups and also no analysis was made to prove that they were equivalent with respect to any external or internal criterion. Each group was assumed to be randomly drawn from the population under consideration.

Table 1 shows the size of each group and some other details.

Instruments

Four tests were used in this study, Vocabulary Test: form A and B, and Mathematics Aptitude Test: form A and B. All items in these four tests were compiled from the Kit of Reference Tests for Cognitive Factors--Revised Edition--(French, Ekstrom, & Price, 1963). The following are details of each test:

Vocabulary Test: Form A (VA). 5 choices, 25 items (12 minutes for all groups).

Example: jovial

1 - refreshing
2 - scare
3 - thickset
4 - wise
5 - jolly

Items in this test are the same as the first 25 items

TABLE 1

NUMBER OF STUDENTS IN EACH GROUP ACCORDING TO SCHOOLS

School System	School	No. of Classes	Group			Total
			CV	CF	EL	
Edmonton Public School	Ottewell J.H.	8	63	61	62	186
	Strathearn El. and J.H.	5	36	41	40	117
Edmonton Separate School	St. Cecilia Sep.	11	87	83	78	248
County of Strathcona	*F.R. Haythorne J.H.	5	40	40	37	117
	Sherwood Hts. J.H.	4	28	29	27	84
County of Parkland	Stony Plain J.H.	4	36	33	37	106
	Spruce Grove J.H.	4	25	31	24	80
St. Albert Separate School	Sir Alexander MacKenzie J.H.	3	31	30	29	90
		44	346	348	334	1028

*One class, not included above, took part in a preliminary testing to set time limit.

in Vocabulary Test--V-2 in the Kit and are also in the same sequence. This test measures verbal comprehension. The test was used under all three test-taking instructions and answered on three different answer sheets. Either one of three instructions were printed on the front cover of a test booklet. Time limits were equal for all groups and set longer than the original test to allow enough time for students to finish their answers.

Vocabulary Test: Form B (VB). 4 choices, 30 items (8 minutes for all groups).

Example: attempt

1 - run
2 - hate
3 - try
4 - stop

Items in this test are the same as the first 30 items in Vocabulary Test--V-1, which is a parallel form of form V-2, in the Kit, and also in the same sequence. This test was used under the Conventional Testing method for all three groups using IBM optical answer sheets. Conventional test instructions were also printed on the front cover of the test booklets. Time limits were set equal to the original form, although the number of items was five less, to allow enough time for students to finish their answers.

Mathematics Aptitude Test: Form A and B (RA-RB). 5 choices, 15 items each (15 minutes for RA and 10 minutes for RB).

Example: How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

1 - 10
2 - 20
3 - 25
4 -100
5 -125

Items in both tests were randomly assigned from all 30 items in Mathematics Aptitude Test--R-1 in the Kit which measures numerical reasoning. Form A was used under three different test-taking instructions and answered on the same answer sheets as used for VA test. Form B was used under the Conventional Testing method and answered on the same IBM answer sheet as VB test. Specific test instructions for each method were also printed on the front cover of the test booklets. The time limit was set longer for RA to allow students enough time to finish answers under the experimental methods, but for RB the time limit was the same as half of the original test.

Answer Sheets. There were three different formats for the answer sheets, one for each test-taking method, for VA and RA tests. The Conventional Testing was answered on an IBM answer sheet and could be scored by machine. The other two forms, especially designed for the students' convenience in answering had to be scored by hand. Answers to VB and RB tests were also on IBM answer sheets and thus scored by machine. It is not impossible, however, to develop answer sheets for these experimental methods for machine scoring but this work should be considered only if these methods proved worth the effort.

Test instructions and answer sheets

are given in Appendix A.

Preliminary Test for Time Limit. Before the first test was taken, a class of 28 grade nine students in F. R. Haythorne Junior High School were given the VA and RA tests under the two experimental test-taking methods, CF and EL. The purpose of this test was to try test-taking instructions and set the time limit. After the testing, oral instructions were revised and time limits were given for each test. Students used for this pilot testing were not included in the main study.

Test-Taking Instructions

There were two types of instructions for VA and RA tests. The first, given orally by the proctors, consisted of general information for all three groups, and covered the way test booklets were distributed, and what students had to do during the testing time. Approximately five minutes was used for this.

The second part was the main and specific instructions for each test-taking method and was printed on the front cover of the test booklets; each booklet had only one set of instructions. These instructions described the method of answering and included examples. They also encouraged the student to ask questions if he was unsure of the correct procedure.

The VB and RB tests had only printed instructions, since this method was already known to all students.

Both types of test-taking instructions are given in Appendix A.

Test Administration Procedures

For both testing times and all 44 classes in this study, a group of graduate and undergraduate students in the Faculty of Education, University of Alberta, were used as test proctors. Most were experienced teachers. Information and details about the project, and the precise procedures to use were given to them before testing. Each proctor was given the "Instructions to Examiners" and "Instructions for Students" when he went testing, and each was responsible for one class at a time. The first instructions were an outline of administration procedures, and the second were the oral instructions to read to the class before testing.

The first testing time began with a distribution of VA test booklets which were pre-arranged in an alternate order: CV, CF and EL test booklets, accompanied by answer sheets. This pre-arranged sequence of test booklets was used for the purpose of random assignment of students into groups. Thus, when the first test was given, all students were then grouped randomly, regardless of their sex or any other criterion. This procedure was convenient and successful in terms of randomization. The proctor, then, read the general instructions to the class, and asked the students to go on reading the specific instructions on their test booklets. A question period was given and followed by students writing the test. The second test, the RA test, was given when the first test was finished. Students received test booklets under the same test-taking instructions. No general instructions were given at this time, and the students were asked to read the specific instructions on the test booklets, and start writing the test after another question period. The RA test was answered on the same answer sheets as used with the VA test.

After the first test session, students' names in each class were recorded according to their groups on separate sheets. These lists were given to proctors at the second testing time. There were two parts at this test session. The first, for VA and RA tests, was carried out in the same manner as that of the first session, except that, at this time, test booklets were distributed to students according to groups and names on the lists. The VB and RB tests were given in the second part at two successive times, and were answered on the same answer sheets. These two tests were answered under the conventional method, so there were no special instructions. Students were asked to read the specific instructions on test booklets and started writing a test after a question period.

A Technique for Differential Weighting of Item Options

The differential weighting procedure used in this study was a simple one if compared with those reported in the literature. Hambleton *et al.* (1970) used two different procedures; one resulted from a rather complex manipulation. The other, which was similar to the technique in this study, was simpler. The second procedure, in their study, started with 22 experts rank-ordering for correctness the five item options. The average rank among these experts was then obtained for each alternative, including the correct answer. The distractor with the second lowest rank was weighted three, and so on to the distractor with the lowest rank which was weighted zero. This procedure resulted in a discrete weighting system for all item options from four, the correct answer, to zero, the least correct distractor, with an equal value of one step apart.

The procedure used in the present study started with 44 raters, a group of undergraduate and graduate students taking courses in Educational Measurement at the University of Alberta during the 1972-73 winter session. These students were asked to rank order the four incorrect options for their degree of correctness, in both VA and RA test items. These ranks were then converted into weights giving four to the highest and so on down to one as the lowest rank. The average weights among all 44 raters for each incorrect option were then obtained with the weight of five assigned to the correct answer. These weights were used to score answers under the conventional test-taking method. Option weights obtained by this technique were not discrete and varied both among options in the same item and across items. This system of weighting seems more reasonable than the second one used by Hambleton et al. (1970), since the differences between option weights reflect their varied degree of correctness. This was the assumption underlying the differential weighting approach. The reliability of rating by the analysis of variance method, using unadjusted reliabilities (Winer, 1970, p. 283), was .895 for VA and .906 for RA tests, suggesting a very high degree of agreement among raters.

Tables of raters by classes, and option weights for each test, VA and RA, are given in Appendix A.

Scoring Techniques

Four different scoring techniques were used in this study. They were applied as follows:

(1) Conventional Scoring (CVS Method). This scoring technique was used to score answers from VA and RA tests under the Conventional Testing and the Elimination Testing, and also all answers from VB and RB tests. Under the Conventional Testing, a correct answer was scored one, otherwise including omission was scored zero. Under the Elimination Testing, an answer with all incorrect alternatives crossed out was scored one, otherwise including omission was scored zero. The total score was a sum of item scores and no guessing formula was used.

(2) Differential Weighted Scoring (DWS Method). This scoring technique was used to score answers from VA and RA tests under the Conventional Testing. Each incorrect option had a predetermined weight obtained from the technique of

differential weighting described above, with the correct answer having a weight of five. An item score was the weight of the option selected, and the total score was a sum of item scores.

(3) Confidence Weighted Scoring (CWS Method). This scoring technique was used to score answers from VA and RA tests under the Confidence Testing. There were five scoring functions used under this scoring technique. In all cases, the total score was a sum of item scores. The following are these five scoring functions:

1. $s_{ij_1} = r_{ij}$
2. $s_{ij_2} = \log r_{ij}$; $s_{ij} = 0$, when $r_{ij} = 0$
3. $s_{ij_3} = 1 - \log |10 - r_{ij}|$; $s_{ij} = 1$, when $r_{ij} = 10$
4. $s_{ij_4} = r_{ij}^2/10$
5. $s_{ij_5} = z_{r_{ij}}$

where: s_{ij} is the item score for the ith person and jth item.

r_{ij} is the degree of confidence (in points) a student assigns to the correct answer (possible values run from 0 to 10).

log is the common logarithm (base 10).

$z_{r_{ij}}$ is a normalized standard score of r_{ij} on the distribution of 11 possible values of r_{ij} (see Table 2 for the method of normalization).

(4) Elimination Scoring (ELS Method). This scoring technique was used to score answers from VA and RA tests under

TABLE 2

NORMALIZED STANDARD SCORE OF r_{ij} ON THE DISTRIBUTION OF 11 POSSIBLE VALUES OF r_{ij}

r_{ij}	f	F	Cumulative Proportion	$z_{r_{ij}}$*
10	1	10.5	0.9545	1.69
9	1	9.5	0.8636	1.10
8	1	8.5	0.7727	0.75
7	1	7.5	0.6818	0.47
6	1	6.5	0.5909	0.23
5	1	5.5	0.5000	0
4	1	4.5	0.4091	-0.23
3	1	3.5	0.3182	-0.47
2	1	2.5	0.2273	-0.75
1	1	1.5	0.1364	-1.10
0	1	0.5	0.0455	-1.69

*Values read from the table of normal distribution.

the Elimination Testing. Each incorrect option crossed out was scored one, and the correct answer, if crossed out, was scored -4. Omissions were scored zero. An item score was an algebraic sum of item scores.

Problems With the Confidence Scoring Functions

Some discussions about scoring functions used in the Confidence Weighted Scoring method are needed to clarify their characteristics. Two important points were considered when using these functions in the study--the problem of the value to be assigned to log 0, and the various possible values of item scores from different scoring functions. The problem of the log 0 value will be discussed first and followed by the problem of item scores.

A Pilot Analysis on the Value of log 0. When Shuford et al. (1966) introduced the reproducible scoring system into the Confidence Testing approach, the logarithmic scoring function, they contended, was the only reproducible scoring function that depended solely on the probability assigned to the correct answer. The function, however, had one difficulty--the unbounded property of log 0 value. They suggested an approximate solution, a truncated logarithmic function setting the value of log 0 at -1. This function was used in studies by Hambleton et al. (1970), and Rippey (1968, 1970), with some minor changes in the form of the function.

In this study, there were two logarithmic functions, one of which was a modified form of that suggested and used in the literature but the other was original. Since the problem of log 0 value could not be avoided, and its value was likely to affect the distribution of test scores and might result in different outcomes for the study, it was contended that the problem should be investigated and determined before any further analysis. A pilot analysis was thus undertaken with answers from VA and RA tests obtained under the Confidence Testing to examine the test reliability. The second scoring function was used in this analysis with three different values of log 0: -1, -.30, and 0. The different sets of possible item scores are shown in Table 3. These score sets differed only on the first value. Reliabilities of test scores (alpha coefficient) from the three sets of possible answers were obtained by the analysis of variance method (Winer, 1970, p. 289). Results of this analysis are shown in Table 4.

The alpha coefficients obtained from the use of three different values of log 0 (Table 4) showed clearly the effect of this problem on test scores. In all cases, the results were consistently in favour of the 0 value. The value of -1, however, gave consistently higher alpha values than those obtained using the value of -.30, except for the second RA test. It is likely that the first possible score set, tying the value of log 0 with log 1, was the best one among them. It was also assumed that the same results would be obtained if applied to the third scoring function.

TABLE 3

THREE SETS OF POSSIBLE ITEM SCORES FROM DIFFERENT VALUES OF log 0

Values of log 0	Possible Item Scores, when r_{ij} =										
	0	1	2	3	4	5	6	7	8	9	10
0	0	0	.30	.48	.60	.70	.78	.85	.90	.95	1.00
-0.30	-0.30	0	.30	.48	.60	.70	.78	.85	.90	.95	1.00
-1.00	-1.00	0	.30	.48	.60	.70	.78	.85	.90	.95	1.00

TABLE 4

RELIABILITIES (ALPHA COEFFICIENTS) OF TEST SCORES OBTAINED FROM THE USE OF DIFFERENT VALUES OF log 0

Values of log 0	VA		RA	
	1st	2nd	1st	2nd
0	.711	.681	.669	.672
-0.30	.702	.660	.665	.658
-1.00	.705	.661	.669	.648

As a result of this pilot study, the value of 0 was given to log 0 in both logarithmic functions for use in the main analysis.

Possible Item Scores from Different Scoring Functions. Scoring functions under Confidence Testing, as shown in the previous section, present one interesting characteristic and, hence, deserve some discussion here. Table 5 shows various possible item scores obtained from their use, and Figure 1 is the plot of these scores (x_m) against their ranks (m). By considering the plotted lines, it is apparent that each function has rates of increase between two successive values that differ from the others. Among these functions, only the first one appears as a straight line; all others are in curves with different forms. This evidence suggests that these functions do not give the same increment between successive possible scores both within and across functions. To see their patterns more clearly, Table 6 was constructed showing differences between two successive possible scores (d_n) obtained under each function. These values (d_n) were then plotted against their ranks (n) as shown in Figure 2. Now, it is clearly seen that what was noted is true. Each scoring function has its typical pattern of possible score increments. Function 1 has a special pattern different from others with all equal score increments resulting in a straight line parallel to the horizontal axis. Other functions have different regular curves. It is noted that the irregularities appearing at the beginning of the second line and at the end

TABLE 5

POSSIBLE ITEM SCORES FROM FIVE SCORING FUNCTIONS

x_m	F1	F2	F3	F4	F5
x_1	0	0	0	0	-1.69
x_2	1	0	.05	.1	-1.10
x_3	2	.30	.10	.4	- .75
x_4	3	.48	.16	.9	- .47
x_5	4	.60	.22	1.6	- .23
x_6	5	.70	.30	2.5	0
x_7	6	.78	.40	3.6	.23
x_8	7	.85	.52	4.9	.47
x_9	8	.90	.70	6.4	.75
x_{10}	9	.95	1.00	8.1	1.10
x_{11}	10	1.00	1.00	10.0	1.69

TABLE 6

DIFFERENCES BETWEEN TWO SUCCESSIVE POSSIBLE ITEM SCORES FROM FIVE SCORING FUNCTIONS

d_n	F1	F2	F3	F4	F5
d_1	1	0	.05	.1	.59
d_2	1	.30	.05	.3	.35
d_3	1	.18	.06	.5	.28
d_4	1	.12	.06	.7	.24
d_5	1	.10	.08	.9	.23
d_6	1	.08	.10	1.1	.23
d_7	1	.07	.12	1.3	.24
d_8	1	.05	.18	1.5	.28
d_9	1	.05	.30	1.7	.35
d_{10}	1	.05	0	1.9	.59

FIGURE 1

A PLOT OF x_m AGAINST m

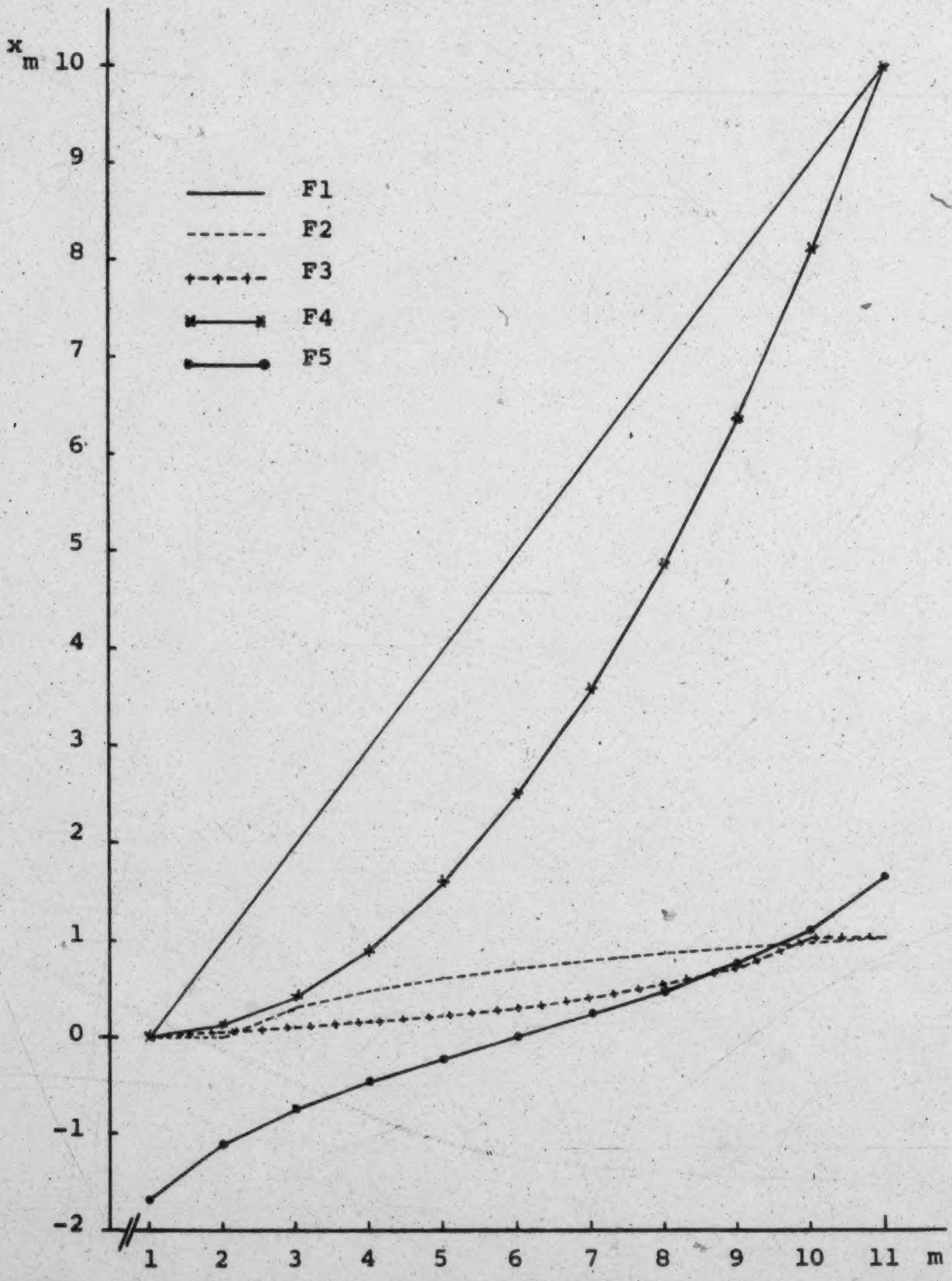
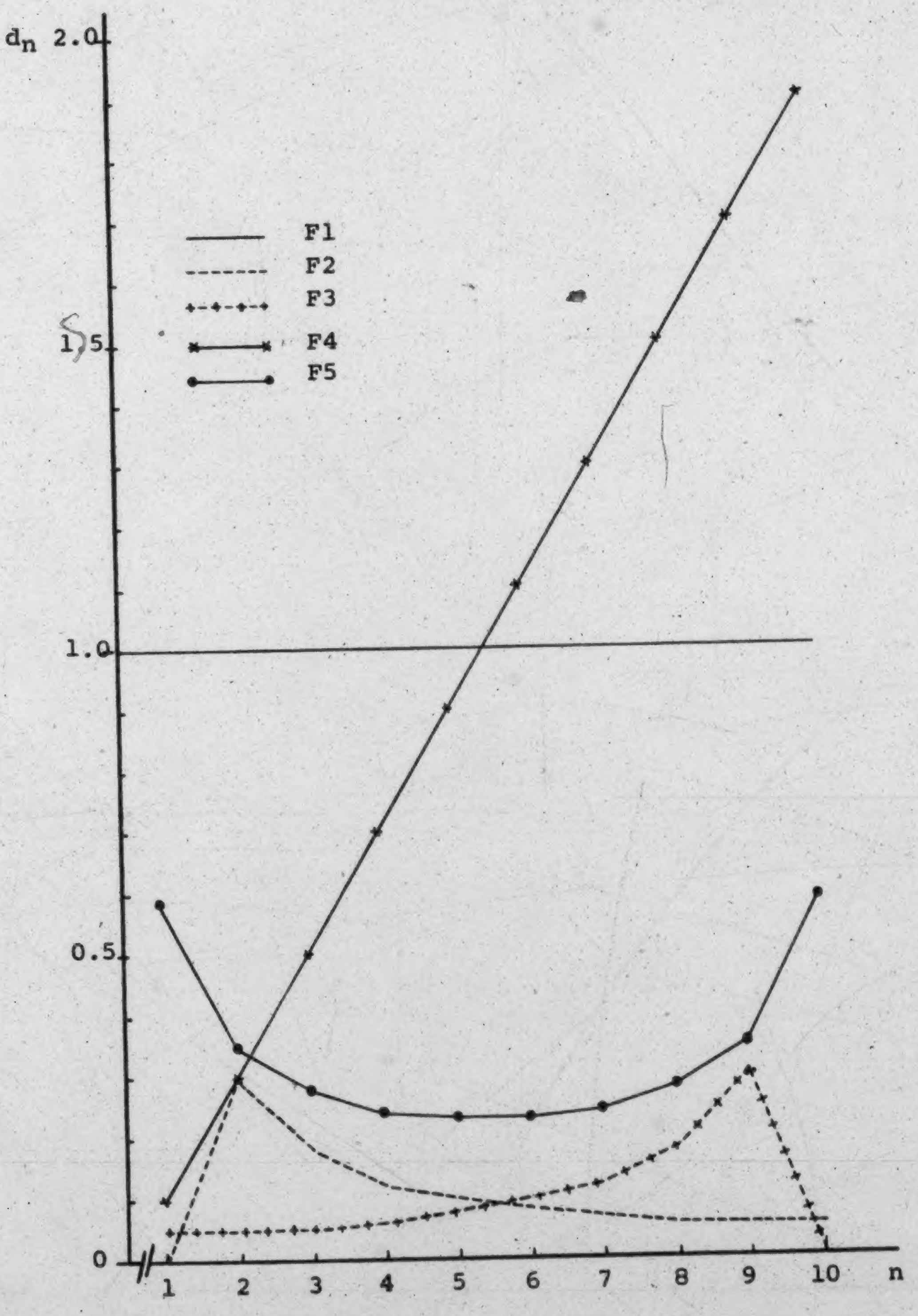


FIGURE 2

A PLOT OF d_n AGAINST n



of the third line in Figure 2 are results of giving the value of 0 to log 0. These irregularities will not be considered here.

The differences among the possible score increments resulting from the use of these functions have significant meaning to test scores and may affect test characteristics. More consideration is needed to understand the underlying meaning and make use of their merits in appropriate ways. The following is a summary of findings about these functions read from Figure 2.

Function 1. This function makes use of the points of confidence assigned to the correct answers without any change. All possible scores are then linearly related to the given confidence levels. When possible points of confidence are in the form of natural numbers from 0 to 10, the increments between successive values are constant, *i.e.*, all equal to one. In terms of assigning scores according to personal confidence, it means that this scoring function gives an equal increment of reward to equal increments of confidence level, no matter how high or low the confidence level is.

Function 2. This function, and all the following, changes possible points of confidence into another series of values. This conventional logarithmic function gives a decreasing increment pattern; the higher the point of confidence is, the smaller is the score increment. In terms of assigning scores according to personal confidence, this scoring function gives a high increment of reward to a low confidence level, and gives a low increment of reward to a higher confidence level. The score increment varies inversely with the confidence level.

Function 3. Functions 3 to 5 were original to this study. Function 3 is logarithmic and was purposely designed to be complementary to the conventional logarithmic one in terms of possible score increments. It gives an increasing increment pattern; the higher the point of confidence is, the larger is the score increment. This function gives more reward to high levels of confidence. The score increment varies directly with personal confidence in the correct answer.

Function 4. This quadratic scoring function behaves in somewhat the same manner as the third one. The difference between them is the pattern of increment. The score increment in this function increases faster than that in the former one.

Function 5. This function is likely to give a compromise between functions 2 and 3, the two logarithmic functions. It gives both decreasing and increasing increments. At the lower part of the confidence distribution, the increment decreases when the confidence level increases, and then reverses at the higher part, with the turning point at the median. This function gives more increment of reward to confidence levels at both extremes.

Consideration of these scoring functions leads one to realize that they represent four distinctive scoring models based on score increments as their main characteristic. Since all scoring functions used in this study with the confidence-weighting test-taking approach depend entirely on the points of confidence assigned to the correct option, these score increment models are also dependent upon this limitation. The following are a presentation of these models in mathematical form:

Let $x_1, x_2, x_3, \ldots, x_m$ be possible item scores from the lowest to the highest values,

$y_1, y_2, y_3, \ldots, y_m$ be equally spaced points of confidence assigned to the correct answer, from the lowest to the highest values, and

$d_n = x_{m+1} - x_m$ be the difference between two successive possible item scores.

Then we have the following four models:

Model 1: Constant Increment Model

$$x_m = f(y_m)$$

and

$$d_n = d_{n+1}$$

This model gives equal distance between two successive possible item scores when the points of confidence increase equally. Function 1 is one scoring function based on this model.

Model 2: Decreasing Increment Model

$$x_m = f(y_m)$$

and

$$d_n > d_{n+1}$$

This model gives decreasing distance between two successive possible item scores when the points of confidence increase equally. Function 2, the conventional logarithmic function, is one form based on this model.

Model 3: Increasing Increment Model

$$x_m = f(y_m)$$

and

$$d_n < d_{n+1}$$

This model is complementary to the second one, and gives increasing distance between two successive possible scores when the points of confidence increase equally. Functions 3 and 4 are among possible forms based on this model.

Model 4: Normalized Increment Model

$$x_m = f(y_m)$$

and

$$x_m \backsim N(0, 1)$$

This model gives a normal distribution of possible item scores when the points of confidence increase equally. Function 5 represents the use of this model.

An innovation in this study is the analysis of the theoretical models underlying the scoring functions. The first two models underlie scoring functions used in previous studies. Models 3 and 4 and the scoring functions based on them are unique to this study.

The analysis of the models was undertaken on the assumption that the different scoring functions would affect test scores in different ways, particularly their reliabilities and validities. These latter factors are the main concern of this study.

School Achievement

School achievement scores used in this study were obtained from three subject areas: Language Arts, Mathematics, and Science. An academic average was also computed based upon four subjects including the above three and Social Science. Marks from schools were obtained in three different forms: letter-grades, stanine-grades, and raw scores. Since marks from the same subjects had to be combined into the same set, each was standardized using the mean and standard deviation of each school group for conversion. A frequency distribution for each school group and each subject was constructed before standardization. The letter-grade system needed special treatment to fit into this table. These grades were first changed into numbers and the numbers were then used to construct the frequency distribution. Table 7 shows the numbers assigned to letter-grades before further treatments.

Four sets of standard scores were then obtained by the procedure described above within each group; thus there were 12 sets of school achievement criteria. These scores were used in the main analysis to provide test validity.

Data Obtained to Test Hypotheses

The design and procedures described in previous

TABLE 7

NUMBERS ASSIGNED TO LETTER-GRADES FOR STANDARDIZING

Grade	Number
A	11
A-	10
B+	9
B	8
B-	7
C+	6
C	5
C-	4
D+	3
D	2
D-	1
F	0

sections were carried on to obtain the following data:

The First Test Session. Eighteen sets of data were gathered from this session, nine for each VA and RA test. They were:

(1) Two sets of scores from each test under the Conventional Testing, one by the Conventional Scoring and the other by the Differential Weighted Scoring.

(2) Five sets of scores from each test under the Confidence Testing using the five scoring functions.

(3) Two sets of scores from each test under the Elimination Testing, one by the Conventional Scoring and the other by the Elimination Scoring.

The Second Test Session. Twenty-four sets of data were gathered at this time. Eighteen of them were the same as obtained from the first session and six other sets were obtained from VB and RB tests, two from each group. Scores from VB and RB tests, though tested and scored under the same methods, were recorded separately by group for the purpose of analysis within group. Thus six score sets were gathered rather than two.

Achievement Criteria. Twelve sets of achievement scores were obtained and recorded in standard score form by the procedure previously described. There were four sets for each group.

Summary of Testing Design and Data Sets

To give an overall view of the testing design and data sets gathered in this study, Tables 8 and 9 were constructed. Table 8 shows the testing and scoring methods for VA and RA tests at both test sessions. As reported previously, there were initially three test-taking methods and four scoring methods used in the study. Table 9 shows all data sets included in the analysis, classified by tests, testing times, and groups. Achievement scores recorded from schools were obtained by conventional examinations so no scoring method was specified.

Statistical Hypotheses

The main purpose of this study was to examine test reliability and validity as indices of the effectiveness of

TABLE 8

TESTING DESIGN FOR VA AND RA TESTS AT BOTH TESTING TIMES

Scoring Method	Test-Taking Method		
	CV	CF	EL
CVS	*	–	*
DWS	*	–	–
CWS_{1-5}	–	*	–
ELS	–	–	*

different testing and scoring methods. The following hypotheses were tested:

(1) The coefficients of internal consistency (alpha coefficient) from test scores under the experimental test-taking and scoring methods should be consistently higher than those obtained under the conventional methods.

(2) The coefficients of stability (test-retest reliability coefficient) obtained from test scores under the experimental methods should be consistently higher than those obtained under the conventional methods.

(3) Test scores obtained under the experimental methods should be more valid, with respect to external criteria, than those obtained under the conventional methods.

Statistical Techniques

Statistical analyses used in this study concerned the following computations:

TABLE 9

A SUMMARY OF DATA SETS IN THE STUDY

Test Session	Test	Group CV	CF	EL	Total
1st Session	VA	CVS DWS	CWS_{1-5}	CVS ELS	9
	RA	CVS DWS	CWS_{1-5}	CVS ELS	9
2nd Session	VA	CVS DWS	CWS_{1-5}	CVS ELS	9
	RA	CVS DWS	CWS_{1-5}	CVS ELS	9
	VB	CVS	CVS	CVS	3
	RB	CVS	CVS	CVS	3
School Achievement	Language Arts	*	*	*	3
	Mathematics	*	*	*	3
	Science	*	*	*	3
	Academic Average	*	*	*	3
	Total	14	26	14	54

*Signifies data obtained from schools.

(1) One-way analysis of variance with repeated measures to obtain internal consistency reliability (alpha coefficient) of test scores, i.e., adjusted reliability (Winer, 1970, p. 289), and reliability of ratings, i.e., unadjusted reliability (Winer, 1970, p. 283).

(2) Product-Moment Correlations, for test-retest reliability and the validity.

(3) Tests of differences between ρ_1 and ρ_2 using independent samples, for validities and test-retest reliabilities between groups (Glass & Stanley, 1970, p. 311).

(4) Tests of differences between ρ_{12} and ρ_{13} using dependent samples, for validities in the same group (Glass & Stanley, 1970, p. 313, and Oklin & Siotani, 1964).

(5) Tests of differences between ρ_{12} and ρ_{34} using dependent samples, for test-retest reliabilities in the same group (Oklin, 1967, p. 113).

(6) Tests of differences between alpha coefficients using independent samples, for alpha coefficients between groups (Feldt, 1969).

CHAPTER IV

RESULTS

The contents of this chapter are grouped in two main sections. The first section deals with the reliability of test scores. Two types of reliability were obtained in this study: internal consistency in terms of the alpha coefficient, and test-retest reliability. These two cases will be treated separately since their characteristics are rather distinctive. The second section deals with the validity of test scores. Two types of criterion scores were used: school achievement based on Language Arts, Mathematics, and Science; and aptitude test scores from vocabulary and mathematics aptitude tests. The first part of the validity section deals with validities of test scores under different scoring methods when school achievement is the criterion. The second part examines validities with respect to aptitude scores. Finally, there is a discussion of the entire validity problem.

In all cases, the results within each group were examined first and then a comparison was made across groups among values chosen as the best from the first stage including those from the conventional method. The test scores and results obtained from the initial test session were used as major indicators of the effectiveness of the experimental techniques. Data from the second test session were used mainly for test-retest reliability and as supporting evidence for the initial results.

The following symbols are used in all tables in this chapter:

VA = Vocabulary Test, Form A
RA = Mathematics Aptitude Test, Form A
VB = Vocabulary Test, Form B
RB = Mathematics Aptitude Test, Form B
CV = Conventional Testing Method
CF = Confidence Testing Method
EL = Elimination Testing Method
CVS_1 = Conventional Scoring Technique under the Conventional Testing Method
CVS_2 = Conventional Scoring Technique under the Elimination Testing Method
DWS = Differential Weighted Scoring Technique under the Conventional Testing Method
CWS_{1-5} = Confidence Weighted Scoring Technique, functions 1 to 5, under the Confidence Testing Method
ELS = Elimination Scoring Technique under the Elimination Testing Method
LA = Language Arts
MA = Mathematics
SC = Science
AV = Academic Average

Reliability

Coefficient of Internal Consistency

Coefficients of internal consistency, i.e., alpha coefficients, were obtained by the analysis of variance method (Winer, 1970, p. 289) since this procedure could be applied to all of the different scoring techniques used in this study. Because there was no procedure available to test the significance of a difference between two alpha values within the same group, the comparison among them was made on the basis of their manifest values and the consistency of the results across tests. A comparison of alpha values across groups was made by a procedure suggested by Feldt (1969). In all cases, critical values for significant difference were those for a two tailed test.

Conventional Testing Group. There were two scoring techniques used under this testing method: Differential Weighted Scoring and Conventional Scoring. Alpha coefficients obtained from test scores under these scoring techniques are shown in Table 10.

The results in Table 10 indicate that alpha coefficients of test scores obtained under the DWS technique were consistently higher than those under the CVS_1 technique. However, since the CVS_1 technique was used as a baseline for all comparisons, values from both the DWS and CVS_1 techniques were retained for further comparisons across groups.

Confidence Testing Group. Five scoring functions,

TABLE 10

ALPHA COEFFICIENTS IN CV GROUP

Test	Scoring Technique	
	DWS	CVS_1
VA_1	.671	.627
VA_2	.739	.673
RA_1	.709	.596
RA_2	.736	.639

TABLE 11

ALPHA COEFFICIENTS IN CF GROUP

Test	Scoring Technique				
	CWS_1	CWS_2	CWS_3	CWS_4	CWS_5
VA_1	.743	.711	.765	.774	.740
VA_2	.722	.681	.759	.771	.718
RA_1	.691	.669	.709	.715	.692
RA_2	.695	.672	.714	.720	.691

labelled CWS_1 to CWS_5, were employed with the tests given to the Confidence Testing group. Alpha coefficients obtained from test scores using these scoring functions are shown in Table 11. Table 12 shows the rank orders of these values from the highest to the lowest within each test.

Table 12 shows that the rank orders of all five scoring functions were very consistent across tests. There was only one exception, RA_1, where alpha values from the CWS_1 and CWS_5 changed their ranks. In all cases the alpha values under the CWS_4 were highest, those under the CWS_3 ranked second, and values from CWS_2 were lowest. Alpha values from the CWS_1 and CWS_5 functions may be considered as ranking third and fourth respectively. Since the alpha values under the CWS_4 and CWS_3 were consistently high relative to all others, these values were retained for further comparisons across groups, _i.e._, there were two scoring techniques chosen from this group for further study.

Elimination Testing Group. There were two scoring techniques used with this testing method: Elimination Scoring and Conventional Scoring. Alpha coefficients obtained from test scores utilizing these techniques are shown in Table 13. The CVS_2 technique gave consistently higher values than did the ELS technique in all tests. However, since the ELS technique is one of the distinctive techniques used in this study, values from both the ELS and CVS_2 techniques were retained for further comparisons across groups.

A Comparison of Alpha Coefficients Across Groups. As

TABLE 12

RANK ORDERS OF ALPHA VALUES FROM THE SAME TEST

Test	Scoring Technique				
	CWS_1	CWS_2	CWS_3	CWS_4	CWS_5
VA_1	3	5	2	1	4
VA_2	3	5	2	1	4
RA_1	4	5	2	1	3
RA_2	3	5	2	1	4

TABLE 13

ALPHA COEFFICIENTS IN EL GROUP

Test	Scoring Technique	
	ELS	CVS_2
VA_1	.647	.741
VA_2	.689	.753
RA_1	.565	.602
RA_2	.640	.670

a result of the comparisons within each group, alpha values from six scoring techniques were retained for comparisons across groups. These were: the DWS and CVS_1 techniques in the CV group; the CWS_4 and CWS_3 in the CF group; and the ELS and CVS_2 techniques in the EL group. Alpha values obtained from these techniques were compared within each test, VA_1 and RA_1, using a test of significant differences between two alpha values suggested by Feldt (1969). The values from VA_2 and RA_2 tests were not considered here since test scores from the second test session were obtained for the purpose of test-retest reliability.

Since the procedure for testing the significance of differences between two alpha values made use of the $\underline{F}$ distribution and the problem was concerned with a two tailed test, the probabilities of the critical values of $\underline{F}$ at .005 and .025 levels were doubled to make the probabilities for the present test .01 and .05 respectively. The degrees of freedom for all tests were between 300 and 350. There was no $\underline{F}$ table that gave degrees of freedom in this neighbourhood. The critical values of $\underline{F}$ used in these tests were calculated for each pair of the degrees of freedom. These values, with their degrees of freedom, are shown in Table 14. When using these critical values, the first degree of freedom is that of the group with the larger value of alpha in the comparison. Results of these tests for VA_1 are shown in Table 15, and for RA_1 in Table 16.

TABLE 14

CRITICAL VALUES OF THE F-RATIO WITH DIFFERENT DEGREES OF FREEDOM FOR TESTS OF DIFFERENCES BETWEEN ALPHA COEFFICIENTS ACROSS GROUPS

Comparison Groups	Degrees of* Freedom	F-ratio at .05 level	F-ratio at .01 level
CV-CF	345,347	1.2350	1.3199
CF-CV	347,345	1.2351	1.3201
CV-EL	345,333	1.2380	1.3242
EL-CV	333,345	1.2375	1.3234
CF-EL	347,333	1.2377	1.3238
EL-CF	333,347	1.2371	1.3228

---

* d.f.1 is always from the group which has a larger value of alpha coefficient in the comparison.

TABLE 15

F-RATIO OF COMPARISONS OF ALPHA COEFFICIENTS FROM VA_1 BETWEEN SCORING TECHNIQUES ACROSS GROUPS

Scoring Techniques	DWS (.67084)	CVS_1 (.62718)	CWS_3 (.76535)	CWS_4 (.77371)	ELS (.64712)	CVS_2 (.74188)
DWS	–	+	1.4028^{**}	1.4546^{**}	1.0721	1.2752^{*}
CVS_1		–	1.5888^{**}	1.6475^{**}	1.0565	1.4444^{**}
CWS_3			–	+	1.5039^{**}	1.1000
CWS_4				–	1.5594^{**}	1.1406^{**}
ELS					–	+
CVS_2						–

TABLE 16

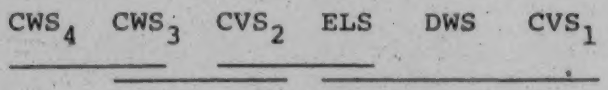
F-RATIO OF COMPARISONS OF ALPHA COEFFICIENTS FROM RA_1 BETWEEN SCORING TECHNIQUES ACROSS GROUPS

Scoring Techniques	DWS (.70893)	CVS_1 (.59590)	CWS_3 (.70857)	CWS_4 (.71493)	ELS (.56462)	CVS_2 (.60152)
DWS	–	+	1.0012	1.0210	1.4958^{**}	1.3690^{**}
CVS_1		–	1.3866^{**}	1.4175^{**}	1.0774	1.0141
CWS_3			–	+	1.4939^{**}	1.3673^{**}
CWS_4				–	1.5273^{**}	1.3978^{**}
ELS					–	+
CVS_2						–

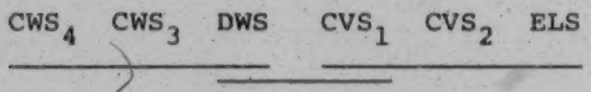
**significant at .01 level. *significant at .05 level.
+ no test available.

To gain a more meaningful picture of the results from these two tables, the information with respect to the significant differences are presented schematically as follows:

VA_1 Test



RA_1 Test



Scoring techniques underlined by a common line do not differ from each other; those not underlined by a common line do differ. In this case, techniques from the same group, e.g., the DWS and CVS_1 are from the CV group, are also underlined by a common line since there was no test for the differences between the values from these techniques.

The results of VA_1 test show that, among six values of alpha coefficient, those from the CWS_4, CWS_3, and CVS_2 are significantly higher than those from the ELS, DWS, and CVS_1. The results of RA_1 test are similar. The alpha values from the CWS_4, CWS_3, and DWS are significantly higher than those from the CVS_1, CVS_2, and ELS. In both tests, the alpha coefficients from the CWS_4 and CWS_3 are higher than others including the values from the typical conventional scoring techniques, the CVS_1. Thus the results from these two tests indicate that the CWS_4 and CWS_3 provide the most reliable test scores in terms of the internal consistency.

Summary of Results

In this section, the focus of the study is on the selection of the scoring techniques which provide test scores with high internal consistency in terms of their alpha coefficients. A comparison among the alpha values was made within each group and then across groups. It was shown that the alpha coefficients from test scores under the CWS_4 and CWS_3 techniques from the CF group were higher than those values under the other techniques including the CVS_1, the typical conventional technique. The analyses show no distinction among the alpha values from the DWS, CVS_1, ELS, and CVS_2 techniques. Thus, the results indicate that the CWS_4 and CWS_3 experimental techniques provide more reliable test scores in terms of the internal consistency than does the conventional technique.

Test-Retest Reliability

Test-retest reliability was measured using a product-moment correlation between test scores from the first and the second test sessions with each scoring technique. Tests of the significance of differences between two *r*'s were available for both dependent and independent cases. A procedure suggested in Oklin's article (1967, p. 113) was used for *r*'s in the same group, and Fisher's *z* transformation and test was used for *r*'s across groups (Glass & Stanley, 1970, p. 311). In all cases, the critical values used were for two tailed tests.

Conventional Testing Group. There were two values of test-retest reliability for each test, one for test scores under the DWS technique and one under the CVS_1 technique. Table 17 shows these values and results of a test of significance of the difference between each pair.

TABLE 17

TEST-RETEST RELIABILITIES IN CV GROUP

Test	Scoring Technique		
	DWS	CVS_1	z
VA	.620	.680	1.0009
RA	.530	.607	1.1646

There were no significant differences between reliabilities under the different scoring techniques. However, the CVS_1 technique provided higher values than did the DWS technique in both tests. These results are different from those obtained using alpha coefficients. For the purpose of a comparison across groups, the reliabilities under the CVS_1 technique were retained for further study.

Confidence Testing Group. There were five values of test-retest reliability for each test resulting from the use of the five scoring functions in this group. Table 18 shows these values, and Tables 19 and 20 show the results of tests of significance of the differences between reliabilities using the different scoring functions.

TABLE 18

TEST-RETEST RELIABILITIES IN CF GROUP

Test	Scoring Technique				
	CWS_1	CWS_2	CWS_3	CWS_4	CWS_5
VA	.764	.715	.792	.797	.753
RA	.669	.648	.684	.688	.667

TABLE 19

RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF VA TEST SCORES (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.764)	CWS_2 (.715)	CWS_3 (.792)	CWS_4 (.797)	CWS_5 (.753)
CWS_1	-	1.0236	0.6965	0.8069	0.2429
CWS_2		-	1.7407	1.9000	0.7772
CWS_3			-	0.1268	0.9199
CWS_4				-	1.0615
CWS_5					-

TABLE 20

RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF RA TEST SCORES (TABLED VALUES ARE z's)

Scoring Techniques	CWS_1 (.669)	CWS_2 (.648)	CWS_3 (.684)	CWS_4 (.688)	CWS_5 (.667)
CWS_1	-	0.3475	0.2591	0.3315	0.0337
CWS_2		-	0.6145	0.6914	0.3134
CWS_3			-	0.0705	0.2942
CWS_4				-	0.3678
CWS_5					-

There were no significant differences among the reliabilities obtained using the different scoring functions for both VA and RA tests. However, since the purpose of this comparison was to choose the one that provided the highest reliability, a rank ordering of these values was used to make the selection. Table 21 shows the results of ranking the reliabilities from the same test.

TABLE 21

RANK ORDERS OF RELIABILITIES WITHIN EACH TEST

Test	Scoring Technique				
	CWS_1	CWS_2	CWS_3	CWS_4	CWS_5
VA	3	5	2	1	4
RA	3	5	2	1	4

The results in Table 21 show perfect agreement between tests. The reliabilities of the CWS_4 ranked first and so on to those of the CWS_2 which ranked last. This result was consistent with that for the alpha coefficients. The reliabilities of test scores under the CWS_4 and CWS_3 techniques were thus retained for further comparisons.

Elimination Testing Group. There were two values of test-retest reliability for each test, one for test scores under the ELS technique and one under the CVS_2 technique. Table 22 shows these values and results of a test of significance of a difference between each pair.

TABLE 22

TEST-RETEST RELIABILITIES IN EL GROUP

Test	Scoring Technique		
	ELS	CVS_2	z
VA	.671	.663	0.1494
RA	.599	.589	0.1469

There were no significant differences between reliabilities under the different scoring techniques. This result was the same as in two previous groups. There was also a consistent indication that one technique, the ELS, provided higher reliabilities than did the other, the CVS_2. Therefore, the reliabilities under the ELS technique were retained for further comparisons. This result was not the same as in the case of the alpha coefficients.

A Comparison of Test-Retest Reliabilities across Groups. As a result of comparisons within each group, reliabilities from test scores under four scoring techniques were retained for this comparison. These were: the CVS_1 technique in the CV group, the CWS_4 and CWS_3 in the CF group, and the ELS technique in the EL group. Reliabilities of test scores under these scoring techniques were compared within each test. A test of significance of the differences between each pair of values was made using Fisher's z transformation and test (see Glass & Stanley, 1970, p. 311). Table 23 shows the results from the VA test and Table 24 from the Ra test.

TABLE 23

RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF VA TEST SCORES ACROSS GROUPS (TABLED VALUES ARE z's)

Scoring Techniques	CVS_1 (.680)	CWS_3 (.792)	CWS_4 (.797)	ELS (.671)
CVS_1	-	3.2394^{**}	3.4230^{**}	0.2206
CWS_3		-	0.1268^{+}	3.4313^{**}
CWS_4			-	3.6132^{**}
ELS				-

**Significant at .01 level.

$+$ z's value obtained from the comparison within group.

Results of the test of differences in the two tables were not quite the same. Reliabilities from the VA test

TABLE 24

RESULTS OF COMPARISONS OF TEST-RETEST RELIABILITIES OF RA TEST SCORES ACROSS GROUPS (TABLED VALUES ARE z's)

Scoring Techniques	CVS_1 (.607)	CWS_3 (.684)	CWS_4 (.688)	ELS (.599)
CVS_1	-	1.7312	1.8361	0.1557
CWS_3		-	0.0705^+	1.8716
CWS_4			-	1.9756^*
ELS				-

*Significant at .05 level.

+z's value obtained from the comparison within group.

scores differed from each other more than did reliabilities from RA test scores. To gain a more meaningful picture of these results, the following summary of information was made:

VA Test

$$\underline{\mathrm{CWS}_4\quad \mathrm{CWS}_3}\quad \underline{\mathrm{CVS}_1\quad \mathrm{ELS}}$$

RA Test

$$\begin{array}{cccc} \mathrm{CWS}_4 & \mathrm{CWS}_3 & \mathrm{CVS}_1 & \mathrm{ELS} \\ \cline{1-3} & & & \\ \cline{2-4} \end{array}$$

In the above schemes, scoring techniques underlined by a common line do not differ from each other, and those not underlined by a common line do differ. It is evident that the reliabilities of test scores under the CWS_4 and CWS_3 techniques were, in general, higher than those under the

CVS_1 and ELS techniques. The reliabilities under the ELS technique were likely to be the lowest. The result that the CWS_4 and CWS_3 techniques provide test scores more reliable than do the other techniques was the same as in the case of the alpha coefficients.

Summary of Results

The purpose of the study in this section was to select the scoring technique which provide test scores with highest test-retest reliability. A comparison among the reliabilities was made within each group and then across groups. The results were consistent with the alpha coefficients. The results show that the CWS_4 and CWS_3 techniques provided more reliable test scores than did the other techniques used in this study. Since the CVS_1 technique was the typical conventional technique, the results also indicate that the CWS_4 and CWS_3 are better techniques in terms of the test-retest reliability than the conventional technique.

Validity

There were two types of criteria used in the study of the validity of test scores. They were: School Achievement and Aptitude Test Scores. It was suspected that test scores under the experimental techniques would correlate differently with different types of criteria. For this reason, the results of the analysis were grouped according to types of criterion scores: the validity with School

Achievement, and with Aptitude Test Scores. Tests of the differences between validities within the same group were done by a test of the differences between two correlations with dependent samples (see Glass & Stanley, 1970, and Oklin & Siotani, 1964). Tests of the differences between validities across groups were done by using the Fisher's z transformation and test (Glass & Stanley, 1970, p. 311). It is noted that the significance of the difference between two validities within the same group depends not only on the two values but also the correlation between test scores from which the validities are obtained. Thus, the same difference does not necessarily mean the same level of significance. In all cases, the critical values used were for two tailed tests.

School Achievement Scores as Criteria

School Achievement scores used in this study were obtained from three subjects: Language Arts, Mathematics, and Science. An academic average was also obtained and used for test validity calculations. There was no evidence to show how reliable these scores were, since they were collected and recorded by schools in terms of total scores. However, a consideration on inter-correlations among these scores within each group suggest that they were reasonable criteria for this study. Tables 25, 26, and 27 show intercorrelations among the four scores within each group.

Results from these tables indicate the same pattern of intercorrelations among the three groups. Tests of the

TABLE 25

INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN CV GROUP

Subjects	LA	MA	SC	AV
LA	-	.585	.604	.799
MA		-	.710	.825
SC			-	.838
AV				-

TABLE 26

INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN CF GROUP

Subjects	LA	MA	SC	AV
LA	-	.595	.608	.795
MA		-	.707	.835
SC			-	.851
AV				-

TABLE 27

INTERCORRELATIONS OF SCHOOL ACHIEVEMENT SCORES IN EL GROUP

Subjects	LA	MA	SC	AV
LA	-	.622	.611	.817
MA		-	.680	.811
SC			-	.820
AV				-

differences among the correlations using the same criterion across groups were also made using Fisher's z transformation and test (Glass & Stanley, 1970, p. 311). No statistically significant differences were obtained. The evidence suggests the comparability of these three groups in terms of school achievement and indicates that these achievement scores could be used as criteria for any of the groups.

The following are the study of the validities of VA and RA test scores under each test-taking method when school achievement scores were used as criteria. Tests of the differences between validities within the same group were done by a test of the difference between two correlations with dependent samples (see Glass & Stanley, 1970, p. 313, and Oklin & Siotani, 1964).

Conventional Testing Group. Table 28 shows the validities of test scores under two scoring techniques, the DWS and the CVS_1. Table 29 shows the results of tests of the differences between validities of the same test under different scoring techniques.

Only two values from VA test scores attained statistical significance, but all were significant in the case of RA test scores. Since all significant differences were in favour of the validities under the CVS_1 technique, it was evident that the validities of test scores under the CVS_1 technique were, in general, higher than those under the DWS technique. Therefore, the validities under the CVS_1 technique were retained for further comparisons across groups.

TABLE 28

VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN CV GROUP

Test	Scoring Technique	Criteria			
		LA	MA	SC	AV
VA_1	DWS	.480	.442	.464	.517
	CVS_1	.486	.445	.472	.517
VA_2	DWS	.438	.387	.449	.491
	CVS_1	.471	.410	.484	.518
RA_1	DWS	.124	.259	.192	.234
	CVS_1	.271	.470	.395	.422
RA_2	DWS	.145	.300	.257	.268
	CVS_1	.296	.514	.461	.475

TABLE 29

DIFFERENCES BETWEEN VALIDITIES UNDER DWS AND CVS_1 SCORING TECHNIQUES (TABLED VALUES ARE z's)

Test	Differences Between Values			
	LA	MA	SC	AV
VA_1	0.4829	0.2358	0.6378	0
VA_2	2.0420*	1.3838	2.1801*	1.7262
RA_1	4.5193**	6.7757**	6.3911**	5.9773**
RA_2	4.6771**	6.9846**	6.5592**	6.6808**

**Significant at .01 level.

*Significant at .05 level.

Confidence Testing Group. Table 30 shows the validities of test scores under five scoring functions, the CWS_1 to CWS_5. Tables 31 to 38 show the results of tests of the differences between validities of test scores from VA, and Tables 39 to 46 from RA, one table for each criterion measure.

Since there were many pairs of values and tables involved, a consideration of the results was made with each test first followed by a general evaluation.

Validities of VA Test Scores

1. The results of the tests of differences in Table 31, when LA was a criterion, were confounded. No summary could be made. But the results in Table 32, for the VA_2 test scores, were clear. The highest validities were obtained for CWS_2, followed by CWS_1 and CWS_5, then CWS_3, and finally CWS_4.

In both tables there were indications that the validities under the CWS_3 were significantly higher than those under the CWS_4.

2. There was only one pair of comparisons that attained statistical significance when MA was a criterion. This was the difference between the validities under the CWS_3 and CWS_4. It is apparent that the validities under all scoring functions in Tables 33 and 34 are generally comparable.

3. The results in Tables 35 and 36 were the same as those in Tables 33 and 34, except that one more pair attained

TABLE 30

VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN CF GROUP

Test	Scoring Technique	Criteria			
		LA	MA	SC	AV
VA_1	CWS_1	.336	.333	.400	.410
	CWS_2	.342	.332	.396	.408
	CWS_3	.323	.332	.397	.406
	CWS_4	.312	.320	.386	.393
	CWS_5	.334	.332	.392	.407
VA_2	CWS_1	.426	.392	.468	.486
	CWS_2	.450	.390	.473	.497
	CWS_3	.388	.383	.453	.462
	CWS_4	.374	.377	.445	.450
	CWS_5	.419	.385	.456	.479
RA_1	CWS_1	.304	.499	.426	.466
	CWS_2	.304	.489	.417	.458
	CWS_3	.300	.502	.433	.469
	CWS_4	.297	.499	.433	.467
	CWS_5	.303	.499	.421	.464
RA_2	CWS_1	.274	.479	.431	.441
	CWS_2	.272	.469	.422	.435
	CWS_3	.272	.483	.435	.442
	CWS_4	.271	.479	.434	.439
	CWS_5	.275	.484	.429	.442

TABLE 31

DIFFERENCES BETWEEN VALIDITIES OF VA$_1$ TEST SCORES WITH LA (TABLED VALUES ARE z's)

Scoring Techniques	CWS$_1$ (.336)	CWS$_2$ (.342)	CWS$_3$ (.323)	CWS$_4$ (.312)	CWS$_5$ (.334)
CWS$_1$	-	0.3385	1.4822	1.9983*	0.4427
CWS$_2$		-	1.0776	1.4476	0.9477
CWS$_3$			-	2.7846**	0.9531
CWS$_4$				-	0.4794
CWS$_5$					-

TABLE 32

DIFFERENCES BETWEEN VALIDITIES OF VA$_2$ TEST SCORES WITH LA (TABLED VALUES ARE z's)

Scoring Techniques	CWS$_1$ (.426)	CWS$_2$ (.450)	CWS$_3$ (.388)	CWS$_4$ (.374)	CWS$_5$ (.419)
CWS$_1$	-	2.3669*	3.8454**	4.1451**	1.0735
CWS$_2$		-	3.4663**	3.6576**	3.4531**
CWS$_3$			-	3.6175**	2.6623**
CWS$_4$				-	3.0380**
CWS$_5$					-

**Significant at .01 level. *Significant at .05 level.

TABLE 33

DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH MA (TABLED VALUES ARE z's)

Scoring Techniques	CWS_1 (.333)	CWS_2 (.332)	CWS_3 (.332)	CWS_4 (.320)	CWS_5 (.332)
CWS_1	-	0.0563	0.1143	1.0846	0.2212
CWS_2		-	0	0.5791	0
CWS_3			-	3.0444**	0
CWS_4				-	0.8083
CWS_5					-

TABLE 34

DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH MA (TABLED VALUES ARE z's)

Scoring Techniques	CWS_1 (.392)	CWS_2 (.390)	CWS_3 (.383)	CWS_4 (.377)	CWS_5 (.385)
CWS_1	-	0.1934	0.9110	1.1989	1.0560
CWS_2		-	0.3879	0.6212	0.5490
CWS_3			-	1.5600	0.1710
CWS_4				-	0.5390
CWS_5					-

**Significant at .01 level.

TABLE 35

DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH SC (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.400)	CWS_2 (.396)	CWS_3 (.397)	CWS_4 (.386)	CWS_5 (.392)
CWS_1	–	0.2317	0.3225	1.2007	1.8134
CWS_2		–	0.0583	0.4963	0.4854
CWS_3			–	2.8639**	0.4456
CWS_4				–	0.4151
CWS_5					–

TABLE 36

DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH SC (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.468)	CWS_2 (.473)	CWS_3 (.453)	CWS_4 (.445)	CWS_5 (.456)
CWS_1	–	0.5046	1.5749	1.9045	1.8765
CWS_2		–	1.1540	1.3922	1.9379
CWS_3			–	2.1480*	0.2708
CWS_4				–	0.7680
CWS_5					–

**Significant at .01 level. *Significant at .05 level.

TABLE 37

DIFFERENCES BETWEEN VALIDITIES OF VA$_1$ TEST SCORES WITH AV (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS$_1$ (.410)	CWS$_2$ (.408)	CWS$_3$ (.406)	CWS$_4$ (.393)	CWS$_5$ (.407)
CWS$_1$	-	0.1165	0.4122	1.4632	0.6854
CWS$_2$		-	0.1173	0.7479	0.1222
CWS$_3$			-	3.3888*	0.0897
CWS$_4$				-	0.9731
CWS$_5$					-

TABLE 38

DIFFERENCES BETWEEN VALIDITIES OF VA$_2$ TEST SCORES WITH AV (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS$_1$ (.486)	CWS$_2$ (.497)	CWS$_3$ (.462)	CWS$_4$ (.450)	CWS$_5$ (.479)
CWS$_1$	-	1.1239	2.5309*	2.9870**	1.1106
CWS$_2$		-	2.0356*	2.3515*	2.0802*
CWS$_3$			-	3.2178**	1.5186
CWS$_4$				-	2.0356*
CWS$_5$					-

**Significant at .01 level. *Significant at .05 level.

TABLE 39

DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH LA (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.304)	CWS_2 (.304)	CWS_3 (.300)	CWS_4 (.297)	CWS_5 (.303)
CWS_1	-	0	0.6188	0.7655	0.2528
CWS_2		-	0.3252	0.4672	0.4378
CWS_3			-	0.9269	0.3390
CWS_4				-	0.5052
CWS_5					-

TABLE 40

DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH LA (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.274)	CWS_2 (.272)	CWS_3 (.272)	CWS_4 (.271)	CWS_5 (.275)
CWS_1	-	0.3066	0.3278	0.3477	0.2505
CWS_2		-	0	0.0677	0.5312
CWS_3			-	0.4334	0.3360
CWS_4				-	0.3470
CWS_5					-

TABLE 41

DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH MA (TABLED VALUES ARE z's)

Scoring Techniques	CWS_1 (.499)	CWS_2 (.489)	CWS_3 (.502)	CWS_4 (.499)	CWS_5 (.499)
CWS_1	-	1.6923	0.5113	0	0
CWS_2		-	1.1605	0.7338	4.6801**
CWS_3			-	1.0211	0.3736
CWS_4				-	0
CWS_5					-

TABLE 42

DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH MA (TABLED VALUES ARE z's)

Scoring Techniques	CWS_1 (.479)	CWS_2 (.469)	CWS_3 (.483)	CWS_4 (.479)	CWS_5 (.484)
CWS_1	-	1.6714	0.7193	0	1.3717
CWS_2		-	1.2340	0.7417	2.8833**
CWS_3			-	1.8965	0.1232
CWS_4				-	0.4767
CWS_5					-

**Significant at the .01 level.

TABLE 43

DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH SC (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.426)	CWS_2 (.417)	CWS_3 (.433)	CWS_4 (.433)	CWS_5 (.421)
CWS_1	-	1.4620	1.1425	0.8078	1.3277
CWS_2		-	1.3695	1.1261	1.8331
CWS_3			-	0	1.4283
CWS_4				-	1.6061
CWS_5					-

TABLE 44

DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH SC (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS_1 (.431)	CWS_2 (.422)	CWS_3 (.435)	CWS_4 (.434)	CWS_5 (.429)
CWS_1	-	1.4650	0.6996	0.3712	0.5335
CWS_2		-	1.1153	0.8665	1.3159
CWS_3			-	0.4631	0.7169
CWS_4				-	0.4630
CWS_5					-

TABLE 45

DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH AV (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (.466)	CWS_2 (.458)	CWS_3 (.469)	CWS_4 (.467)	CWS_5 (.464)
CWS_1	-	1.3289	0.5007	0.1180	0.5440
CWS_2		-	0.9627	0.6474	2.8003**
CWS_3			-	0.6673	0.6093
CWS_4				-	0.2726
CWS_5					-

TABLE 46

DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH AV (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (.441)	CWS_2 (.435)	CWS_3 (.442)	CWS_4 (.439)	CWS_5 (.442)
CWS_1	-	0.9840	0.1758	0.2485	0.2684
CWS_2		-	0.6042	0.2904	1.3249
CWS_3			-	1.3918	0
CWS_4				-	0.2792
CWS_5					-

** Significant at .01 level.

statistical significance. With SC as a criterion, the differences between the validities under the CWS_3 and CWS_4 were clearly in favour of the CWS_3. All other values were comparable.

4. The results in Table 37, when AV was used as a criterion, showed only one pair attaining statistical significance. This was the difference between validities under the CWS_3 and CWS_4. The results in Table 38 were clearer than those in Table 37. The validity under the CWS_2 was significantly higher than all except the CWS_1 function. The validity under the CWS_4 seemed to be the lowest. The difference between validities under the CWS_3 and CWS_4 also attained statistical significance.

The results of tests of the differences between validities of test scores from VA were generally inconsistent, except for the difference between those values under the CWS_3 and CWS_4 which favoured the CWS_3. No scoring function provided consistently high validities when school achievement scores were used as the criteria.

Validities of RA test scores

The results of tests of the differences between validities of RA test scores are readily examined. Among all pairs of comparisons, there were only three which attained statistical significance. These were the differences between validities under the CWS_2 and CWS_5 in Tables 41, 42 and 45.

It is apparent that no scoring function provided higher validity than did the others. The only exception was a tendency for the validities under the CWS_5 to be higher than those under the CWS_2, especially in the case where MA was a criterion.

Since there was no scoring technique showing its superiority over the others in terms of validity when school achievement was the criterion, the validities under these scoring techniques were retained for comparisons with those validities from the other groups.

Elimination Testing Group. Table 47 shows the validities of test scores under two scoring techniques, the ELS and CVS_2, and Table 48 shows the results of tests of the differences between validities of test scores under these scoring techniques.

The results of the comparisons between validities of test scores for VA are apparent. All pairs attained statistical significance consistently showing the superiority of the ELS technique over the CVS_2 technique. The results in the case of RA test scores were, however, less clear but still consistent in the case of RA_2. All significant differences in this case were also in favour of the ELS technique. It is likely that these two scoring techniques affected VA test scores more than RA test scores. For the purpose of a comparison across groups, the validities of test scores under the ELS technique were retained for further study.

TABLE 47

VALIDITIES OF TEST SCORES WITH SCHOOL ACHIEVEMENT IN EL GROUP

Test	Scoring Technique	Criteria			
		LA	MA	SC	AV
VA_1	ELS	.395	.350	.432	.435
	CVS_2	.254	.214	.324	.285
VA_2	ELS	.382	.326	.431	.430
	CVS_2	.245	.175	.273	.249
RA_1	ELS	.355	.556	.405	.462
	CVS_2	.325	.533	.385	.432
RA_2	ELS	.344	.499	.395	.452
	CVS_2	.292	.463	.351	.394

TABLE 48

DIFFERENCES BETWEEN VALIDITIES UNDER ELS AND CVS_2 SCORING TECHNIQUES (TABLED VALUES ARE _z_'s)

Test	Differences Between Values			
	LA	MA	SC	AV
VA_1	4.0368**	3.8382**	3.1625**	4.3531**
VA_2	4.4973**	4.8701**	5.2495**	5.9798**
RA_1	1.6572	1.4290	1.1323	1.7433
RA_2	2.7999**	2.0994*	2.4218*	3.3580**

** Significant at .01 level. *Significant at.05 level.

VB and RB Test Scores as Criteria

VB and RB tests were given to all students at the second test session. These tests were administered and scored under the conventional method. Therefore, the alpha reliabilities were computed from the total group. These were .778 for VB and .563 for RB. Table 49 shows the correlations of test scores from these two tests with School Achievement within each group.

TABLE 49

CORRELATIONS OF VB AND RB TEST SCORES WITH SCHOOL ACHIEVEMENT

Test	School Achievement			
	LA	MA	SC	AV
CV-VB	.470	.438	.442	.505
CF-VB	.394	.352	.447	.476
EL-VB	.471	.347	.421	.455
CV-RB	.235	.382	.348	.368
CF-RB	.192	.319	.331	.306
EL-RB	.203	.316	.253	.242

Tests of the differences between the correlations of the same test with each criterion were made using Fisher's z transformation and test (Glass & Stanley, 1970, p. 311). No difference attained statistical significance. This suggests the comparability of VB and RB test scores as criteria among the three groups.

Details are given below of the study of the validities of VA and RA test scores under each test-taking method when VB and RB test scores were used as criteria. Tests of the differences between validities within the same group were done by a test of the difference between two correlations with dependent samples (see Glass & Stanley, 1970; p. 313, and Oklin & Siotani, 1964).

Conventional Testing Group. Table 50 shows the validities of test scores under two scoring techniques, the DWS and CVS_1, and Table 51 shows the results of tests of the differences between validities of the same test under different scoring techniques.

The results of tests of the differences between the validities of test scores from RA are clear. All validities under the CVS_1 were significantly higher than those under the DWS. However, the results from the VA test are not as definite. There was only one difference that attained statistical significance in favour of the CVS_1. For the purpose of a comparison across groups, the validities of test scores from both tests under the CVS_1 technique were retained. This selection of the CVS_1 technique was the same as that when using School Achievement as the criterion.

Confidence Testing Group. Table 52 shows the validities of test scores under five scoring functions, the CWS_1 to CWS_5. Tables 53 to 56 show the results of tests of the differences between the validities of test scores from VA, and Tables 57 to 60 from RA, one table for each criterion.

TABLE 50

VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN CV GROUP

Test	Scoring Technique	Criteria	
		VB	RB
VA_1	DWS	.584	.312
	CVS_1	.589	.302
VA_2	DWS	.639	.360
	CVS_1	.670	.376
RA_1	DWS	.133	.390
	CVS_1	.277	.496
RA_2	DWS	.244	.481
	CVS_1	.383	.562

TABLE 51

DIFFERENCES BETWEEN VALIDITIES UNDER DWS AND CVS_1 SCORING TECHNIQUES (TABLED VALUES ARE z's)

Test	Criteria	
	VB	RB
VA_1	0.4359	0.7392
VA_2	2.2630*	0.9497
RA_1	4.4328**	3.5618**
RA_2	4.4145**	2.8832**

**Significant at .01 level. *Significant at .05 level.

TABLE 52

VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN CF GROUP

Test	Scoring Technique	Criteria	
		VB	RB
VA_1	CWS_1	.646	.329
	CWS_2	.638	.338
	CWS_3	.636	.307
	CWS_4	.627	.295
	CWS_5	.648	.328
VA_2	CWS_1	.673	.335
	CWS_2	.672	.318
	CWS_3	.647	.314
	CWS_4	.630	.313
	CWS_5	.672	.312
RA_1	CWS_1	.429	.488
	CWS_2	.429	.475
	CWS_3	.426	.493
	CWS_4	.422	.494
	CWS_5	.427	.483
RA_2	CWS_1	.393	.530
	CWS_2	.390	.512
	CWS_3	.390	.537
	CWS_4	.386	.539
	CWS_5	.394	.522

TABLE 53

DIFFERENCES BETWEEN VALIDITIES OF VA$_1$ TEST SCORES WITH VB (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS$_1$ (.646)	CWS$_2$ (.638)	CWS$_3$ (.636)	CWS$_4$ (.627)	CWS$_5$ (.648)
CWS$_1$	-	0.5586	1.4021	1.9389	0.5472
CWS$_2$		-	0.1400	0.6552	1.4538
CWS$_3$			-	0.9346	1.2826
CWS$_4$				-	1.7380
CWS$_5$					-

TABLE 54

DIFFERENCES BETWEEN VALIDITIES OF VA$_2$ TEST SCORES WITH VB (TABLED VALUES ARE *z*'s)

Scoring Techniques	CWS$_1$ (.673)	CWS$_2$ (.672)	CWS$_3$ (.647)	CWS$_4$ (.630)	CWS$_5$ (.672)
CWS$_1$	-	0.1207	3.1799**	4.0793**	0.1882
CWS$_2$		-	1.7084	2.4497*	0
CWS$_3$			-	5.0766**	2.6033**
CWS$_4$				-	3.3993**
CWS$_5$					-

**Significant at .01 level. *Significant at .05 level.

TABLE 55

DIFFERENCES BETWEEN VALIDITIES OF VA_1 TEST SCORES WITH RB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (. 329)	CWS_2 (.338)	CWS_3 (.307)	CWS_4 (.295)	CWS_5 (.328)
CWS_1	-	0.5067	2.4940*	2.8151**	0.2208
CWS_2		-	1.7505	2.0657*	1.1821
CWS_3			-	1.9814*	1.8107
CWS_4				-	2.2081*
CWS_5					-

TABLE 56

DIFFERENCES BETWEEN VALIDITIES OF VA_2 TEST SCORES WITH RB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (.335)	CWS_2 (.318)	CWS_3 (.314)	CWS_4 (.313)	CWS_5 (.312)
CWS_1	-	1.5989	2.0634*	1.7144	3.3641**
CWS_2		-	0.2153	0.2323	0.6398
CWS_3			-	0.2536	0.1661
CWS_4				-	0.0655
CWS_5					-

**Significant at .01 level. *Significant at .05 level.

TABLE 57

DIFFERENCES BETWEEN VALIDITIES OF RA_1 TEST SCORES WITH VB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (.429)	CWS_2 (.429)	CWS_3 (.426)	CWS_4 (.422)	CWS_5 (.427)
CWS_1	-	0	0.4895	0.8070	0.5329
CWS_2		-	0.2575	0.4931	0.9225
CWS_3			-	1.3012	0.1192
CWS_4				-	0.4439
CWS_5					-

TABLE 58

DIFFERENCES BETWEEN VALIDITIES OF RA_2 TEST SCORES WITH VB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS_1 (.393)	CWS_2 (.390)	CWS_3 (.390)	CWS_4 (.386)	CWS_5 (.394)
CWS_1	-	0.4809	0.5141	0.8474	0.2620
CWS_2		-	0	0.2832	0.7404
CWS_3			-	1.8066	0.4685
CWS_4				-	0.7254
CWS_5					-

TABLE 59

DIFFERENCES BETWEEN VALIDITIES OF RA$_1$ TEST SCORES WITH RB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS$_1$ (.488)	CWS$_2$ (.475)	CWS$_3$ (.493)	CWS$_4$ (.494)	CWS$_5$ (.483)
CWS$_1$	-	2.1786*	0.8461	0.7188	1.3751
CWS$_2$		-	1.5930	1.3838	1.5490
CWS$_3$			-	0.3391	1.2335
CWS$_4$				-	1.0132
CWS$_5$					-

TABLE 60

DIFFERENCES BETWEEN VALIDITIES OF RA$_2$ TEST SCORES WITH RB (TABLED VALUES ARE $\underline{z}$'s)

Scoring Techniques	CWS$_1$ (.530)	CWS$_2$ (.512)	CWS$_3$ (.537)	CWS$_4$ (.539)	CWS$_5$ (.522)
CWS$_1$	-	3.0782**	1.3031	1.1868	2.2523*
CWS$_2$		-	2.2675*	2.0657*	1.9816*
CWS$_3$			-	0.9886	1.8991
CWS$_4$				-	1.6718
CWS$_5$					-

**Significant at .01 level. *Significant at .05 level.

The following is a summary of the results from Tables 53-60:

1. The results of the tests on the validities of VA_1 test scores with VB, Table 53, show that all values are comparable. The results in the case of VA_2 test scores, Table 54, give an indication that the validities under the CWS_1 and CWS_5 are a little better than others. However, there was no evidence to suggest that any one was the best among all.

2. The results of the tests on the validities of VA_{1-2} test scores with RB, Tables 55-56, were unclear, and confounded in the case of VA_2 test scores, Table 56. There was no evidence for a single best technique.

3. There was no difference in Tables 57 and 58 attaining statistical significance. All values were comparable.

4. There was an indication, in Tables 59-60, that the validities under the CWS_1 were somewhat superior to the others in the case of the validities of RA test scores with RB, and that results under the CWS_2 were likely to be the lowest. However, no selection for the best one could be made.

Since there was no evidence from these results to support the selection of one scoring technique over the others, the validities obtained under all scoring techniques were retained for a comparison across groups in the next part of this section.

Elimination Testing Group. Table 61 shows the

TABLE 61

VALIDITIES OF TEST SCORES WITH VB AND RB TEST SCORES IN EL GROUP

Test	Scoring Technique	Criteria	
		VB	RB
VA_1	ELS	.672	.285
	CVS_2	.546	.236
VA_2	ELS	.664	.272
	CVS_2	.583	.224
RA_1	ELS	.384	.383
	CVS_2	.361	.387
RA_2	ELS	.381	.360
	CVS_2	.354	.369

TABLE 62

DIFFERENCES BETWEEN VALIDITIES UNDER ELS AND CVS_2 SCORING TECHNIQUES (TABLED VALUES ARE _z_'s)

Test	Criteria	
	VB	RB
VA_1	4.3263**	1.3709
VA_2	3.2626**	1.5442
RA_1	1.2884	0.2255
RA_2	1.6487	0.4949

**Significant at .01 level.

validities of test scores under two scoring techniques, the ELS and CVS_2, and Table 62 the results of tests of the difference between values under these different techniques.

There were only two pairs of differences that attained statistical significance. Both were from the validities of VA test scores with VB and were in favour of the ELS technique. All other values were comparable. It is apparent that the scoring techniques affected the validities of VA test scores more than those of RA. However, for the purpose of a comparison across groups, the validities of test scores under the ELS technique were retained for further study. This selection was the same as in the case of using School Achievement scores as criteria.

A Comparison of Validities of Test Scores Across Groups. As the results of a comparison within each group with one type of criterion at a time, the same scoring techniques were retained from the CV and EL groups. These were: the CVS_1 and ELS. The comparisons in the CF group also yielded the same results for both types of criteria and all scoring techniques provided comparable values of validity. Therefore, the validities from all scoring techniques used in this group were retained for a comparison across groups. Therefore, there were seven values of validity for each test with the same criterion, and six criterion scores involved in this part of the study. Table 63 shows the validities of VA_1 and RA_1 test scores with all criteria. VA_2 and RA_2 test scores were not considered in this part

TABLE 63

VALIDITIES OF VA_1 and RA_1 TEST SCORES UNDER SEVEN SCORING TECHNIQUES

Test	Scoring Techniques	Achievement Criteria				Aptitude Criteria	
		LA	MA	SC	AV	VB	RB
VA_1	CVS_1	.486	.445	.472	.517	.589	.302
	CWS_1	.336	.333	.400	.410	.646	.329
	CWS_2	.342	.332	.396	.408	.638	.338
	CWS_3	.323	.332	.397	.406	.636	.307
	CWS_4	.312	.320	.386	.393	.627	.295
	CWS_5	.334	.332	.392	.407	.648	.328
	ELS	.395	.350	.432	.435	.672	.285
RA_1	CVS_1	.271	.470	.395	.422	.277	.496
	CWS_1	.304	.499	.426	.466	.429	.488
	CWS_2	.304	.489	.417	.458	.429	.475
	CWS_3	.300	.502	.433	.469	.426	.493
	CWS_4	.297	.499	.433	.467	.422	.494
	CWS_5	.303	.499	.421	.464	.427	.483
	ELS	.355	.556	.405	.462	.384	.383

since test scores from the second test session were used mainly for test-retest reliability.

Because the tests of differences between validities used in this part made use of the critical values from the unit normal distribution and the sizes of these groups were very comparable (346 for the CV, 348 for the CF, and 334 for the EL groups), it was possible to compare these values to see the sequences of their ranks with each criterion. This arrangement was used to give an indication of whether the comparisons among them should be made separately with each type of criterion or should be combined. Table 64 was thus constructed using rank orders of all values with the same criterion. The rank orders of the validities from the CF group was the same because they were assumed comparable as a result of the study within this group.

The results in Table 64 show a consistent pattern of rank orders across criteria within each type. When School Achievement scores were criteria, the consistent pattern held across criteria, but when Aptitude Test Scores (VB and RB) were criteria, the consistent pattern held across test scores. This evidence suggests that different types of criteria affected test validities and that an examination should be made within each type to gain a better view of the results. Tables 65 and 66 show the results of the examination when school achievement scores served as criteria, and Tables 67 and 68 when VB and RB test scores were criteria. Tests of the differences between these validities across

TABLE 64

RANK ORDERS OF THE VALIDITIES WITH THE SAME CRITERION

Test	Scoring Techniques	Achievement Criteria				Aptitude Criteria	
		LA	MA	SC	AV	VB	RB
VA_1	CVS_1	1	1	1	1	3	1
	CWS_{1-5}	3	3	3	3	2	
	ELS	2	2	2	2	1	3
RA_1	CVS_1	3	3	3	3	3	1
	CWS_{1-5}	2	2	1	1	1	2
	ELS	1	1	2		2	3

TABLE 65

COMPARISONS OF THE VALIDITIES OF VA_1 TEST SCORES WHEN SCHOOL ACHIEVEMENT SCORES WERE CRITERIA (TABLED VALUES ARE *z*'s)

Scoring Techniques	Criteria			
	LA	MA	SC	AV
CVS_1 - CWS_1	2.3869*	1.7312	1.1541	1.7836
CVS_1 - CWS_2	2.2820*	1.7443	1.2197	1.8230
CVS_1 - CWS_3	2.5705*	1.7443	1.2066	1.8492
CVS_1 - CWS_4	2.7279**	1.9148	1.3771	2.0459*
CVS_1 - CWS_5	2.4131*	1.7443	1.2853	1.8361
CVS_1 - ELS	1.4666	1.4666	0.6489	1.3757
CWS_1 - ELS	-0.8968	-0.2469	-0.4939	-0.1300
CWS_2 - ELS	-0.7928	-0.2599	-0.5589	-0.1690
CWS_3 - ELS	-1.0788	-0.2509	-0.5489	-0.1950
CWS_4 - ELS	-1.2347	-0.4289	-0.7148	-0.3899
CWS_5 - ELS	-0.9228	-0.2599	-0.6262	-0.1820

**Significant at .01 level. *Significant at .05 level.

TABLE 66

COMPARISONS OF THE VALIDITIES OF RA_1 TEST SCORES WHEN SCHOOL ACHIEVEMENT SCORES WERE CRITERIA (TABLED VALUES ARE z's)

Scoring Techniques	Criteria			
	LA	MA	SC	AV
CVS_1 - CWS_1	-0.4721	-0.4984	-0.4852	-0.7213
CVS_1 - CWS_2	-0.4721	-0.3279	-0.3410	-0.5902
CVS_1 - CWS_3	-0.4197	-0.5377	-0.6033	-0.7738
CVS_1 - CWS_4	-0.3672	-0.4984	-0.6033	-0.7344
CVS_1 - CWS_5	-0.4590	-0.4984	-0.4066	-0.6951
CVS_1 - ELS	-1.2070	-1.5185	-0.1557	-0.6360
CWS_1 - ELS	-0.7408	-1.0268	0.3249	0.0780
CWS_2 - ELS	-0.7408	-1.1957	0.1820	-0.0519
CWS_3 - ELS	-0.7928	-0.9878	0.4419	0.1300
CWS_4 - ELS	-0.8448	-1.0268	0.4419	0.0910
CWS_5 - ELS	-0.7538	-1.0268	0.2469	0.0520

TABLE 67

COMPARISONS OF THE VALIDITIES OF VA_1 TEST SCORES WHEN VB AND RB WERE CRITERIA (TABLED VALUES ARE z's)

Scoring Techniques	Criteria	
	VB	RB
$CVS_1 - CWS_1$	-1.2197	-0.3934
$CVS_1 - CWS_2$	-1.0492	-0.5246
$CVS_1 - CWS_3$	-0.9967	-0.0656
$CVS_1 - CWS_4$	-0.7738	0.1049
$CVS_1 - CWS_5$	-1.2590	-0.3672
CVS_1 - ELS	-1.7911	0.2466
CWS_1 - ELS	0.5849	0.6369
CWS_2 - ELS	0.7538	0.7668
CWS_3 - ELS	0.8058	0.3119
CWS_4 - ELS	1.0268	0.1430
CWS_5 - ELS	0.5459	0.6109

TABLE 68

COMPARISONS OF THE VALIDITIES OF RA_1 TEST SCORES WHEN VB AND RB WERE CRITERIA (TABLED VALUES ARE z's)

Scoring Techniques	Criteria	
	VB	RB
$CVS_1 - CWS_1$	-2.2951*	0.1312
$CVS_1 - CWS_2$	-2.2951*	0.3541
$CVS_1 - CWS_3$	-3.5541**	0.0393
$CVS_1 - CWS_4$	-2.1771*	0.0262
$CVS_1 - CWS_5$	-2.2557*	0.2098
$CVS_1 - ELS$	-1.5704	1.8170
$CWS_1 - ELS$	0.7019	1.6896
$CWS_2 - ELS$	0.7019	1.4687
$CWS_3 - ELS$	1.9496	1.7806
$CWS_4 - ELS$	0.5849	1.7936
$CWS_5 - ELS$	0.6629	1.6117

**Significant at .01 level. *Significant at .05 level.

groups were done by using the Fisher's z transformation and test (see Glass & Stanley, 1970, p. 311).

Validities With School Achievement. Table 65 shows the results of tests of the differences between the validities of VA_1 test scores, Table 66 of RA_1 test scores.

1. The results shown in Table 64 indicate that the validities of VA_1 test scores under the CVS_1 technique consistently ranked first, those under the ELS technique, second, and under the CWS_{1-5}, last. But, it was observed in Table 65 that, among six pairs of the comparisons that attained statistical significance, five of them were the differences between the validities under the CVS_1 and CWS_{1-5} techniques with LA, favouring the CVS_1. However, the tendency for the validities under the CVS_1 technique to be higher than those under the CWS_{1-5} and ELS techniques was apparent throughout the first part of Table 65. It was evident that the validities of VA test scores under the CVS_1 technique were consistently higher than those under other scoring techniques when School Achievement Scores were criteria, especially when LA was used.

2. Table 64 shows that the results were reversed in the case of RA test scores. The validities under the ELS technique ranked first, with those under the CWS_{1-5}, second, and under the CVS_1, last, with only one exception in the case of SC criterion. However, Table 66 shows that no difference attained statistical significance. It was evident that, though the validities under the ELS technique

were consistently higher than those under the CVS_1 technique, the differences were not large enough for a conclusion in favour of the ELS technique.

Validities with Aptitude Test Scores. Table 67 shows the results of tests of the differences between the validities of VA_1 test scores, and Table 68 of RA_1 test scores.

1. The results in Table 64 indicate that, when Aptitude Test Scores were criteria, the patterns of rank order of the validities were not the same across criteria, but were similar across tests. When VB was a criterion, the validities under the CVS_1 technique ranked third in both VA and RA tests, but those under the CWS_{1-5} and ELS techniques interchanged their ranks. When RB was a criterion, the validities under the ELS technique ranked third, and those under the CVS_1 were likely to rank the first, though it was not clear in the case of VA with RB.

2. Considering the results of tests of the differences in Tables 67 and 68, only the differences between the validities of RA test scores under the CVS_1 and CWS_{1-5} with VB criterion attained statistical significance. However, a tendency toward similar results was also seen in the case of the VA test with the VB criterion. This evidence indicates that, when VB test was a criterion, the validities of test scores under the CVS_1 technique were lower than those under the CWS_{1-5} and ELS. However, no selection between the CWS_{1-5} and ELS techniques could be made since no difference

between those values attained statistical significance.

3. The differences between the validities when RB was a criterion were not statistically significant. Thus, no method provided consistently higher validities than any other. There was an apparent tendency for the validities under the ELS technique to be the lowest.

Summary of Results

In this section, the focus of the study was on the selection of the scoring techniques which would provide highest validity when related to school achievement and aptitude. The results indicated that the scoring techniques affected test validity in different ways. For some tests and some criteria, the conventional scoring technique, *i.e.*, the CVS_1 technique, provided higher validities than did the others; for example, in the case of VA test with School Achievement. But for some other cases, the ELS and CWS_{1-5} techniques provided higher validities than did the conventional one; for example, in the case of RA test with VB test. Although there were some consistent patterns of these values, the patterns did not hold over tests or types of criteria, and, in most cases, results were not statistically significant. Thus, there was no indication that the experimental techniques improve the validities of tests when compared with the conventional scoring procedure.

CHAPTER V

CONCLUSIONS

Summary

The purpose of this study was to compare several methods of scoring multiple-choice tests in terms of the reliability and validity of test scores. Subjects were 1028 grade nine students randomly assigned to three comparably sized groups according to test-taking method. They were: the conventional, the confidence, and the elimination methods. There were four scoring techniques used under these testing procedures. They were: conventional scoring, differential weighted scoring, confidence weighted scoring, and elimination scoring. Conventional scoring was used with the conventional and elimination methods; differential weighted scoring with the conventional method; confidence weighted scoring, which had five scoring functions, with the confidence method; and elimination scoring with the elimination method. Two aptitude tests were used with these experimental methods. They were a vocabulary and a mathematics aptitude test. Two types of criterion scores were obtained: school achievement scores, and aptitude test scores from similar forms of the vocabulary and mathematics aptitude tests.

There were two testing sessions. During the first, two aptitude tests were taken under the experimental methods. On the second, the same two tests were repeated and two other aptitude tests were administered to obtain test scores for use as criteria. The aptitude tests as criteria were administered and scored with conventional procedure. The school achievement scores were obtained from schools at the end of the first term.

The analyses of data were designed to obtain three statistics for test scores under each scoring technique. They were: the internal consistency of test scores in terms of alpha coefficient, the test-retest reliability, and the validity with school achievement, and aptitude test scores. Tests of the differences between values in each type of test statistic were made. A selection of the best method was made according to the results of these tests, where possible.

Findings and Implications

The ultimate goal of this study was to examine differences in test reliability and validity using the new test-taking and scoring methods, employing as a baseline, results obtained by administering and scoring tests under the conventional procedure, *i.e.*, conventional scoring with conventional testing. Test-taking and scoring methods used in this study, except the conventional one, have been proposed for their ability to assess students' partial knowledge and eliminate guessing. The study was designed to examine the following queries:

1. Whether test scores obtained under any of the experimental methods are more reliable than those obtained under the conventional method (evidence examined included both the internal consistency of test scores and the stability of test scores over a period of time), and

2. Whether test scores obtained under any of the experimental methods have higher validity for outside criteria measuring either (a) the same or similar traits or (b) school achievement, than those obtained under the conventional method.

The analysis carried out and presented in Chapter four revealed that test scores obtained from two scoring functions, i.e., functions four and three, under the confidence testing were more reliable than those obtained with the conventional method. The evidence is apparent in both the internal consistency of test scores in terms of the alpha coefficient, and test-retest reliability. The analysis indicated, however, that no experimental method provided test scores more valid for outside criteria than did the conventional one. Thus, the hypothesis regarding higher reliability of confidence weighted scores was confirmed, but no evidence supported the hypothesis regarding higher validity.

All scoring techniques used in the present study, except three scoring functions under the confidence testing, i.e., functions three, four and five, were previously studied. A summary of the results of some former studies is given in Table 69. There are only two studies that showed significant gains over conventional testing and scoring, one in terms of reliability and one, validity. The others revealed both increased and decreased values of the two test characteristics. The evidence from these studies indicates no agreement regarding the superiority of either experimental method over the conventional method.

In this study, among the three techniques previously used (thus excluding functions three, four and five under the confidence testing), the differential weighted scoring (DWS) is the only one technique that resulted in more reliable test scores than does the conventional technique (CVS_1), and only in terms of the internal consistency (see Table 10, Chapter IV). None of them provides test scores more valid than does the conventional one. The confidence weighted scoring techniques referred to in this case, are the CWS_1 and CWS_2 in this study. Thus, the results of the present study on these three scoring techniques previously used, combined with those former results can lead to a fairly definite conclusion. The three scoring methods, i.e., the DWS, ELS, and the CWS with the techniques as specified in Table 69, cannot be expected to improve both the reliability and validity of test scores. Further study on these scoring techniques seems unwarranted.

Three new scoring functions used under the confidence testing method are introduced in this study. These are:

TABLE 69

RESULTS FROM SOME PREVIOUS STUDIES COMPARING EXPERIMENTAL METHODS WITH CONVENTIONAL METHOD

Scoring Method	Technique	Authors	Experimental Tests	Reliability		Validity	
				Type	Results	Criterion Measures	Results
Differential Weighted Scoring (DWS)	Experts' judgement	Nedelsky (1954)	Achievement	?	Increased	-	-
	Modified empirical weights	Davis & Fifer (1959)	Aptitude	Parallel forms	Increased*	Teachers' ratings and Aptitude	Decreased
	Empirical weights	Sabers & White (1969)	Aptitude	Split-half	Increased	Achievement	Increased
	Modified experts' judgement	Hambleton et. al. (1970)	Achievement	Split-half	Decreased+	Achievement	Increased
	Experts' average rank	Hambleton et. al. (1970)	Achievement	Split-half	Increased+	Achievement	Increased
	Modified empirical weights	Collet (1971)	Aptitude	Parallel forms	Decreased	Achievement	Decreased
	Modified empirical weights	Hendrickson (1971)	Aptitude	Internal consistency	Increased+	Aptitude	not clear
Elimination Scoring (ELS)	Free Choice	Dressel & Schmid (1953)	Achievement	K-R 20	Decreased+	-	-
	Incorrect option selection	Coombs et. al. (1956)	Aptitude	K-R 20	Increased+	-	-
	Incorrect option selection	Collet (1971)	Aptitude	Parallel forms	Increased	Achievement	Increased*
Confidence Weighted Scoring (CWS)	Degree-of-Certainty	Dressel & Schmid (1953)	Achievement	K-R 20	Increased+	-	-
	10 points of confidence	Michael (1968)	Achievement	Split-half	Increased+	-	-
	100 points of confidence	Hambleton et. al. (1970)	Achievement	Split-half	Decreased+	Achievement	Increased
	H-M-L levels of confidence	Hopkins et. al. (1973)	Achievement	Split-half	Increased	Achievement	Decreased

* Significantly higher than the values from conventional method.

+ No tests made.

(1) CWS_3, the logarithmic function complementary to that used in previous studies (see Hambleton et al., 1970); (2) CWS_4, the quadratic function; and (3) CWS_5, the normalized scoring function (see Chapter III). The CWS_3 and CWS_4 functions are based on the increasing increment scoring model, and the CWS_5 function, the normalized increment scoring model (see Chapter III). These two scoring models have been presented for the first time. The results of the analysis show that the CWS_4 and CWS_3 under the confidence testing condition provide test scores more reliable than does the conventional method, both in terms of the internal consistency and test-retest reliability. None of these three functions, in addition to those previously used, results in more valid test scores than conventional scores. In fact, for some criteria, the validities obtained from these functions are significantly lower than those from the conventional method (see Table 65, Chapter IV). Though the reliabilities from the CWS_4 and CWS_3 functions under the confidence testing method are higher than those from the conventional method, this fact does not necessarily point to the superiority of these techniques since there is no increased validity from either experimental technique. The use of one or the other of these two scoring functions in normal testing practice may, therefore, not be worth the effort.

There are two important studies reported in the literature in which the confidence testing approach is compared with the conventional method. These are the studies by Hambleton _et al_. (1970) and Hopkins _et al_. (1973). These studies are concerned with both reliability and validity. Hambleton _et al_. (1970) used a logarithmic function (which corresponds to the CWS_2 function used in the present study) with 100 points of confidence. They found that the use of confidence weighted scoring resulted in decreased reliability but increased validity. Since the increment in test validity did not attain statistical significance, the authors did not take the result as definitive. They pointed out that three weak points contributed to the insignificant increment: (1) the small group size (211 for all three groups), (2) the longer testing time required for the confidence testing, and (3) the too low difficulty level of test for the subjects employed.

The results of the study by Hambleton _et al_. (1970) led to another study by Hopkins _et al_. (1973). These authors postulated that:

> . . . If the increase in reliability is the result of a gambling response style, it is conceivable that validity could actually decrease even though reliability is increased (Hopkins _et al_., 1973, p. 138).

Hopkins _et al_. (1973) instructed students to place an H, M, or L beside their response, indicating high, medium, or low confidence in the answer given. No scoring formula was used. The item scores depended on the level of confidence given and on the correctness of the responses. They found that the use of such a technique resulted in increased reliability but decreased validity. No difference attained statistical significance. The conclusions from these results are that:

> . . . the added reliable variance often observed in confidence-weighting studies may be irrelevant response style variance and does not increase validity, in fact, it may actually diminish validity (Hopkins et al., 1973, p. 140).

The results of the present study in the case of the CWS_4 and CWS_3 functions under the confidence testing condition tend to support this conclusion. The use of these two scoring functions resulted in increased reliability but decreased validity. Although the scoring techniques used in the two studies are not identical, it is apparent that the results of the present study confirm those results found by Hopkins et al. (1973).

The fact that an increase in test reliability can be accompanied by a decrease in validity is not new. It is well known that, when we are trying to make test scores more reliable in terms of internal consistency, we are increasing the homogeneity of test items or, in this case, test scores. It is also known that increased heterogeneity of a test leads to increased opportunity for raising the validity (Magnusson, 1967, pp. 179-194). Therefore, when we are trying to make test scores more reliable in terms of internal consistency, i.e., the alpha coefficient, and valid at the same time, we may be undertaking an impossible task. However, this task may be accomplished if we at the same time make the criterion more homogeneous and ensure that it measures exactly the same factor as the experimental test does. In this study, the criteria were not modified in any way. The school achievement scores were to a large extent multi-factor measurements. Because we had multi-factor criteria and more homogeneous test scores, i.e., test scores from the CWS_4 and CWS_3 functions, we could not expect a higher correlation between them than when test scores were less homogeneous, i.e., test scores from the conventional method.

The aptitude test criteria, on the other hand, were originally constructed to measure specific factors. Items in these tests were more homogeneous than those in the school achievement tests. But, because we did not try to improve their homogeneity as we did with the experimental tests, aptitude test scores as criteria, i.e., test scores from VB and RB tests, were relatively less homogeneous than the experimental test scores, i.e., test scores from VA and RA tests under the CWS_4 and CWS_3 functions. In this case it could be expected that the validity of the more reliable test scores, i.e., test scores from the CWS_4 and CWS_3 functions, would not be much higher than those of the less reliable test scores, i.e., test scores from the conventional method, when these aptitude test scores, i.e., test scores from VB and RB tests, were criteria.

Evidence to support the above discussion is seen in

Table 70. In this table, the validities of the vocabulary test (VA) with Language Arts (LA) and another vocabulary test (VB) are shown. It is apparent that, when school achievement was criterion, the validity of test scores under the conventional method was significantly higher than those under the confidence method. When an aptitude test was criterion, the results reversed, although no differences attained statistical significance. The results shown in Table 71, for the mathematics aptitude test (RA) are not as clear. The validity with Mathematics (MA) under the conventional method tended to be lower than others, and when another mathematics aptitude test (RB) was the criterion, all values were comparable. Since the differences of the validities with LA did not attain statistical significance, the results neither contradict nor support the above conclusions. The case of the mathematics aptitude test (RA) is likely to be a result of the test's nature, since it is evident that, when a comparison among groups was made, both in the case of test-retest reliability and the validity (Chapter III), the differences between values from the mathematics aptitude test (RA) were less marked than those from the vocabulary test (VA). This is likely an indication that the test-taking and scoring methods have less effect on the mathematics aptitude test than on the vocabulary test.

The ultimate goal of the present study, as well as

TABLE 70

COMPARISONS OF THE VALIDITIES OF VA TEST SCORES UNDER THE CONVENTIONAL AND CONFIDENCE METHODS (VALUES ARE z's)

Test	Criteria	
	LA	VB
VA - CVS_1	.486*	.589
VA - CWS_3	.323	.636
VA - CWS_4	.312	.627

*Significantly higher than the other two values with LA.

TABLE 71

COMPARISONS OF THE VALIDITIES OF RA TEST SCORES UNDER THE CONVENTIONAL AND CONFIDENCE METHODS (VALUES ARE z's)

Test	Criteria	
	MA	RB
RA - CVS_1	.470	.496
RA - CWS_3	.502	.493
RA - CWS_4	.499	.494

of the others reviewed in Chapter II, was to find new test-taking and scoring techniques that can eliminate two crucial disadvantages inherent in the use of the conventional method. These are: (1) the inability to assess partial knowledge and (2) the encouragement of guessing. It has been suggested by several test specialists that the test-taking and scoring techniques used in this study can eliminate these two disadvantages (Coombs et al., 1956; de Finnetti, 1965; Hopkins et al., 1973; Shuford et al., 1966; Wang & Stanley, 1970). But, since the new techniques do not appear to result in increased validity of test scores-which is the most important characteristic of a test--the use of these techniques in practical testing may not be worth the effort. It is apparent that, in order to overcome the disadvantages of the conventional technique, one should look for new approaches rather than pursuing these methods.

Suggestions for Further Research

Because this study was confined to one specific level of subjects and only two aptitude tests, a study on the same problem with other levels of subjects and other tests is recommended. The results of such a study combined with the present one will, no doubt, make the findings more meaningful and the implications more comprehensive.

In this study, no effort was made to find the influence of nonintellectual factors on subjects' test scores. Since there is possibility that subjects' performance is affected by the influences of these factors, there should be a study on this problem. The results of such a study may lead to improvements of the techniques used in the present investigation.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aiken, L. R. Effect on test score variance of differential weighting of item responses. *Psychological Reports*, 1967, *21*, 585-590.

________. Weighting and guessing on varieties of the multiple-choice item. *Educational and Psychological Measurement*, 1968, *28*, 1087-1101.

________. Scoring for partial knowledge on the generalized arrangement item. *Educational and Psychological Measurement*, 1970, *30*, 87-94.

Collet, L. S. Elimination scoring: an empirical evaluation. *Journal of Educational Measurement*, 1971, *8*, 209-214.

Coombs, C. H., J. E. Milholland and F. B. Womer. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, *16*, 13-37.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

Davis, F. B., and G. Fifer. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, *19*, 159-170.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple-choice test items. *Educational and Psychological Measurement*, 1959, *19*, 291-298.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 87-123.

Dressel, P. L., and J. Schmid. Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 1953, *13*, 574-595.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, *16*, 407-424.

________. Confidence weighting and test reliability. *Journal of Educational Measurement*, 1965, *2*, 49-57.

Ebel, R. L. Review of Shuford and Massengill "Valid confidence testing-demonstration kit." Journal of Educational Measurement, 1968, 5, 353-354.

Echternacht, G. J., R. F. Boldt, and W. S. Sellman. Personality influences on confidence test scores. Journal of Educational Measurement, 1972, 9, 235-241.

Echternacht, G. J. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 1969, 34, 363-373.

French, J. W., R. B. Ekstrom, and I. A. Price. Kit of reference tests for cognitive factors. Princeton: Educational Testing Service, 1963.

Glass, G. V., and J. C. Stanley. Statistical Methods in Education and Psychology. Englewood Cliffs: Prentice-Hall Inc., 1970.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of Mental Tests. New York: John Wiley & Sons, Inc., 1950.

Guttman, L., and I. M. Schlesinger. Systematic construction of distractors for ability and achievement test items. Educational and Psychological Measurement, 1967, 27, 569-580.

Hambleton, R. K., D. M. Roberts, and R. E. Traub. A comparison of the reliability and validity of two methods of assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 1970, 7, 75-82.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. Journal of Educational Measurement, 1971, 8, 9-14.

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. Journal of Educational Measurement, 1971, 8, 291-296.

Henver, K. A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359-362.

Hopkins, K. D., A. R. Hakstian, and B. R. Hopkins. Validity and reliability consequences of confidence weighting, *Educational and Psychological Measurement*, 1973, *33*, 135-141.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, *6*, 153-160.

Lord, F. M., and M. R. Novick. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley, 1968.

Magnusson, D. *Test Theory*. Reading: Addison-Wesley, 1967.

Michael, J. J. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, *5*, 307-314.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 1954, *14*, 459-472.

Oklin, I. Correlations revisited. In Stanley, J. (ed.) *Improving Experimental Design and Statistical Analysis*. Chicago: Rand McNally, 1967.

Oklin, I., and M. Siotani. Asympotic distribution functions of a correlation matrix. *Report No. 6*, Stanford, California: Stanford University Laboratory for Quantitative Research in Education, 1964.

Rippey, R. Probabilistic testing. *Journal of Educational Measurement*, 1968, *5*, 211-215.

_______. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970, *7*, 165-170.

Sabers, D. L., and G. W. White. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 1969, *6*, 93-96.

Shuford, E. H., A. Albert, and H. E. Massengill. Admissible probability measurement procedures. *Psychometrika*, 1966, *31*, 125-145.

Soderquist, H. O. A new method of weighting scores in a true-false test. *Journal of Educational Research*, 1936, *30*, 290-292.

Stanley, J. C. Reliability. In Thorndike, R. L. (ed.) *Educational Measurement*. Washington, D.C.: American Council on Education, 1971.

Stanley, J. C., and M. D. Wang. Weighting test items and test-item options, an overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970, 30, 21-35.

Stokes, R. R. The split-response technique. Phi Delta Kappen. 1966, 47, 271-272.

Thorndike, R. L., and E. Hagen. Measurement and Evaluation in Psychology and Education. New York: John Wiley & Sons, Inc., 1969.

Thorndike, R. L. (ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Wang, M. W., and J. C. Stanley. Differential weighting: a review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-706.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.

APPENDIX A

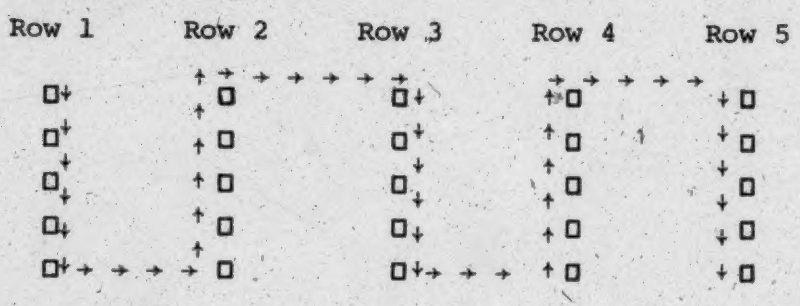
TEST INSTRUCTIONS, ANSWER SHEETS, AND OPTION WEIGHTS

INSTRUCTIONS TO EXAMINERS

1. Tell the students that these tests are part of a study on improving methods for marking multiple-choice exams. Emphasize that it is important that they answer all questions both quickly and accurately.

2. Have the students clear their desks of everything except a pencil. (If no pencil is available, a pen is OK as a last resort).

3. Distribute the answer sheets and the VA tests according to the scheme outlined below (or your version of it).



4. Have the students put their names on the answer sheets and under Faculty or School write the name of their school and their room number.

5. Read the general instructions to the class, have them read the instructions on their test booklets, and then check the class for difficulties in understanding the instructions.

6. Begin and time the test. Total time = 12 min.

7. Have one student pick up the VA tests while you distribute the RA tests according to the same sequence as used previously.

8. Have the students read the instructions on the RA booklets.

9. Begin and time the test. Total time = 15 min.

10. Collect the RA tests and answer sheets, thank the students and teacher, and leave.

INSTRUCTIONS FOR STUDENTS

THESE INSTRUCTIONS ARE TO BE READ BY THE EXAMINER AFTER THE STUDENTS HAVE BEEN GIVEN THE VA TEST BOOKLETS AND ANSWER SHEETS

1. The multiple-choice tests which you are going to write today will be answered in three different ways. One method is the same as you usually use, but the other two are different. Each of you will use only one method.
2. The instructions for answering the test are already printed on the cover of each test booklet. Read these instructions carefully and make sure that you understand them before you start writing. If after reading the instructions, you have any questions, please ask for help. I will explain the method to you.
3. I will give you about 5 minutes to read the test instructions and ask questions. Do not open the test booklet until I tell you to do so. Everyone will start the test at the same time.
4. Now all of you have the test booklets and answer sheets--begin reading the instructions. Be sure you understand the instructions before you start writing the test.

WHEN ALL STUDENTS HAVE READ THE INSTRUCTIONS AND CLEARED UP ANY DIFFICULTIES, CONTINUE AS FOLLOWS:

5. Is everyone ready? Begin!

12 MINUTES LATER:

6. Close your booklets please!

REPEAT STEPS 4, 5, and 6 FOR THE SECOND TEST. REMEMBER THE SECOND TEST (RA) IS 15 MINUTES LONG.

VOCABULARY TEST -- VA-CV.

This is a test of your knowledge of word meanings. Look at the example below. One of the five numbered words has the same meaning or nearly the same meaning as the word above the numbered words. In this example the right answer has already been marked. This was done by placing a black mark between the guidelines on the answer sheet as shown, by using an HB pencil.

i. jovial

1-refreshing
2-scare
3-thickset
4-wise
5-jolly

i. 1 2 3 4 5

To mark an answer, decide first which is the best answer. Then, on the answer sheet, find the row of the answer numbered the same as the question. Make a black mark between the guidelines for the best answer. Make only one mark for each question.

Your score will be the number of the questions correctly answered.

You will have 12 minutes to answer all 25 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

VOCABULARY TEST -- VA-CF.

This is a test of your knowledge of word meanings. There are 25 questions to be answered. Each question has one word given above and 5 numbered words given below. One of them has the same meaning or nearly the same meaning as the word above and is the correct answer. Your task is not to determine the correct answer, but to indicate your confidence in the correctness of each available word. You have 10 points of confidence to distribute among the 5 words. If you are sure that one word, say the 1st. word, is correct, you may give 10 points to that word and give 0 to all others. But if you are not sure about any of them, you may give only a major portion of 10 points to one word which you have more confidence in than the others, and then distribute the rest of the 10 points to some other words.

Look at the example below. In this example, the 5 th. word, jolly, is the correct word, so all other words are incorrect.

i. jovial

1-refreshing
2-scare
3-thickset
4-wise
5-jolly

A. If you are sure that the 5th. word, jolly, is correct, you may give 10 points of confidence to that word by writing '10' under the corresponding number on the answer sheet, and then write '0' under all other numbers, as shown below -

i.	1	2	3	4	5
	0	0	0	0	10

B. If you think that the 5th. word is correct, but you are not very sure and you still think that the 2nd. word, scare, might be correct, you may give 7 points to the 5th. and 3 points to the 2nd. words by writing '7' and '3' under the corresponding numbers on the answer sheet and giving other numbers '0', as shown below -

i.	1	2	3	4	5
	0	3	0	0	7

or, you may give 6 points to the 5th. word, 2 to the 2nd. and the 3rd. words on the answer sheet as in this example -

i.	1	2	3	4	5
	0	2	2	0	6

TURN TO THE NEXT PAGE

or, give them as in this example -

1.	1	2	3	4	5
	0	4	1	0	5

You are free to give any number of points to each answer in an item, but you have to make sure that the points you give to all words add up to 10; no more; no less!

Remember that you have to answer all questions on the answer sheet by writing numbers of points under the corresponding number of words, as shown in the examples.

Your score for an item will be the number of points you give to the correct answer. If you give 10 to the correct answer you receive a score of 10, if you give 7 you receive 7, and if you give 0 you receive 0. So your score for an item will vary from 0 to 10.

It is important for you to know that you will receive a higher score on the test if you indicate honestly your degree of confidence in the correctness of each choice.

You will have 12 minutes to answer all 25 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

VOCABULARY TEST -- VA-EL.

This is a test of your knowledge of word meanings. There are 25 questions to be answered. Each question has one word given above and fivenumbered words given below. One of the five numbered words has the same meaning or nearly the same meaning as the word above and is the correct answer. Your task is not to determine the correct answer, but to determine, among the five numbered words, which of them are incorrect.

Actually, there are 4 numbered words which are incorrect. You should choose only words which you are sure are incorrect. It is not necessary that you find all 4 incorrect words. You may choose only one or two or three words which you are sure are wrong.

Look at the example below. In this example, the 5th. word, jolly, is the correct answer, so all other words are incorrect. If you can find all of them as in this example, then, on the answer sheet on the row of the answer numbered the same as the question, you cross out all numbers except number 5.

i. jovial

1-refreshing
2-scare
3-thickset
4-wise
5-jolly

i. ~~1~~ ~~2~~ ~~3~~ ~~4~~ 5

Your score for an item will be the number of incorrect words you cross out. You will score 4 points from the answer shown above. If, by mistake, you also cross out the correct word, you will lose 4 points for that word.

Suppose that you cross out numbers2,3,4 and 5 on your answer sheet for the above question. You will score 3 points from words numbered 2,3 and 4, but lose 4 points for number 5, so your score for this question would be -1. But if you cross out only numbers 2,3 and 4, your score will be 3 instead of -1 or 4.

So, be careful to cross out only words which you are really sure that they are incorrect. DO NOT GUESS.

You will have 12 minutes to answer all 25 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

MATHEMATICS APTITUDE TEST -- RA-CV.

In this test you will be asked to solve some problems in mathematics. Solve each problem and mark your answer on the answer sheet as shown by the following example -

i. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

1-10
2-20
3-25
4-100
5-125

i. 1 2 3 4 5

To mark your answer, decide first which is the right answer. Then, on the answer sheet, find the row of the answer numbered the same as the question. Make a black mark, with an HB pencil, between the guidelines for the right answer. Make only one mark for each question.

Your score will be the number of the questions correctly answered.

You will have 15 minutes to answer all 15 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

MATHEMATICS APTITUDE TEST -- RA-CF.

In this test you will be asked to solve some problems in mathematics. There are 15 problems to be solved. Each problem has 5 numbered answers. One of them is correct, the other 4 are incorrect. Your task is not to determine the correct answer, but to indicate your confidence in the correctness of each available answer. You have 10 points of confidence to distribute among the 5 answers. If you are sure that one answer, say the 1st. answer, is correct, you may give 10 points to that answer and give 0 to all others. But if you are not sure about any of them, you may give only a major portion of 10 points to one answer which you have more confidence in than the others, and then distribute the rest of the 10 points to some other answers.

Look at the example below. In this example, the 2nd. answer, 20, is the correct answer, so all others are incorrect.

i. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

1-10
2-20
3-25
4-100
5-125

A. If you are sure that the 2nd. answer is correct, you may give 10 points of confidence to that answer by writing '10' under the corresponding number on the answer sheet, and then write '0' under all other numbers, as shown below -

	1	2	3	4	5
i.	0	10	0	0	0

B. If you think that the 2nd. answer is correct, but you are not very sure and you still think that the 3rd. answer, 25, might be correct, you may give 7 points to the 2nd. and 3 points to the 3rd. answers by writing '7' and '3' under the corresponding numbers on the answer sheet and giving other numbers '0', as shown below -

	1	2	3	4	5
i.	0	7	3	0	0

or, you may give 6 points to the 2nd. answer, 2 to the 3rd. and the 4th. answers on the answer sheet as in this example -

	1	2	3	4	5
i.	0	6	2	2	0

TURN TO THE NEXT PAGE

You are free to give any number of points to each answer in an item, but you have to make sure that the points you give to all answers add up to 10; no more; no less.

Remember that you have to answer all questions on the answer sheet by writing numbers of points under the corresponding number of answers, as shown in the examples.

Your score for an item will be the number of points you give to the correct answer. If you give 10 to the correct answer you receive a score of 10, if you give 7 you receive 7, and if you give 0 you receive 0. So your score for an item will vary from 0 to 10.

It is important for you to know that you will receive a higher score on the test if you indicate honestly your degree of confidence in the correctness of each choice.

You will have 15 minutes to answer all 15 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

MATHEMATICS APTITUDE TEST -- RA-EL.

In this test you will be asked to solve some problems in mathematics. There are 15 problems to be solved. Each problem has five numbered answers. One of them is correct, the other four are incorrect. Your task is not to determine the correct answer, but to determine, among the five answers, which of them are incorrect.

Actually there are 4 numbered answers which are incorrect. You should choose answers which you are sure are incorrect. It is not necessary that you find all 4 incorrect answers. You may choose only one or two or three answers which you are sure are wrong.

Look at the example below. In this example, the 2nd. answer, 20, is the correct answer, so all others are incorrect. If you can find all of them as in this example, then, on the answer sheet on the row of the answer numbered the same as the question, you cross out all numbers except number 2.

1. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?
   1-10
   2-20
   3-25
   4-100
   5-125

1. ~~1~~ 2 ~~3~~ ~~4~~ ~~5~~

Your score for an item will be the number of incorrect answers you cross out. You will score 4 points from the answer shown above. If, by mistake, you also cross out the correct answer, you will lose 4 points for that answer.

Suppose that you cross out numbers 1,2,3 and 4, you will score 3 points for numbers 1,3 and 4, but lose 4 points for crossing out number 2, so your score for this question is -1. But if you cross out only numbers 1,3 and 4, your score will be 3 instaed of -1 or 4.

So, be careful to cross out only answers which you are really sure that they are incorrect. DO NOT GUESS.

You will have 15 minutes to answer all 15 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

VOCABULARY TEST -- VB.

This is a test of your knowledge of word meanings. Look at the example below. One of the four numbered words has the same meaning or nearly the same meaning as the word above the numbered words. In this example, the right answer has already been marked. This was done by placing a black mark between the guidelines on the answer sheet as shown, by using an HB pencil.

i. attempt

1-run
2-hate
3-try
4-stop

i. 1 2 3 4
==== ==== ▬▬ ====

To mark an answer, decide first which is the best answer. Then, on the answer sheet, find the row of the answer numbered the same as the question. Make a black mark between the guidelines for the best answer. Make only one mark for each question.

Your score will be the number of the questions correctly answered

You will have 8 minutes to answer all 30 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

MATHEMATICS APTITUDE TEST -- RB.

In this test you will be asked to solve some problems in mathematics. Solve each problem and mark your answer on the answer sheet as shown by the following example -

1. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

   1-10
   2-20
   3-25
   4-100
   5-125

	1	2	3	4	5
1.	:::::	▬	:::::	:::::	:::::

To mark an answer, decide first which is the right answer. Then, on the answer sheet, find the row of the answer numbered the same as the question. Make a black mark, with an HB pencil, between the guidelines for the right answer. Make only one mark for each question.

Your score will be the number of the questions correctly answered.

You will have 10 minutes to answer all 15 questions in this test.

DO NOT MARK IN THIS TEST BOOKLET. If you have any questions, please ask us now.

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

NAME ______________________________
Last First Middle

FACULTY or SCHOOL ______________________________

AGE ______ Years YEAR or GRADE ______ ☐ Male ☐ Female DATE ______ Day Month Year

I.D. NUMBER 147

0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9

1. VA-CV TEST

2. RA-CV TEST

Indicate response by placing a mark between the guidelines as shown in the example. Use HB pencil. Don't make marks longer than guidelines.

Example
1 2 3 4 5

Vocabulary Test -- VA-CV.

1. 1 2 3 4 5
2. 1 2 3 4 5
3. 1 2 3 4 5
4. 1 2 3 4 5
5. 1 2 3 4 5
6. 1 2 3 4 5
7. 1 2 3 4 5
8. 1 2 3 4 5
9. 1 2 3 4 5
10. 1 2 3 4 5
11. 1 2 3 4 5
12. 1 2 3 4 5
13. 1 2 3 4 5
14. 1 2 3 4 5
15. 1 2 3 4 5
16. 1 2 3 4 5
17. 1 2 3 4 5
18. 1 2 3 4 5
19. 1 2 3 4 5
20. 1 2 3 4 5
21. 1 2 3 4 5
22. 1 2 3 4 5
23. 1 2 3 4 5
24. 1 2 3 4 5
25. 1 2 3 4 5
26. 1 2 3 4 5
27. 1 2 3 4 5
28. 1 2 3 4 5
29. 1 2 3 4 5
30. 1 2 3 4 5

Please keep this answer sheet for one more test.

Mathematics Aptitude Test -- RA-CV.

1. 1 2 3 4 5
2. 1 2 3 4 5
3. 1 2 3 4 5
4. 1 2 3 4 5
5. 1 2 3 4 5
6. 1 2 3 4 5
7. 1 2 3 4 5
8. 1 2 3 4 5
9. 1 2 3 4 5
10. 1 2 3 4 5
11. 1 2 3 4 5
12. 1 2 3 4 5
13. 1 2 3 4 5
14. 1 2 3 4 5
15. 1 2 3 4 5

Please place this answer sheet inside the test booklet and hand them in.

CAUTION — AVOID PLACING ANY MARKS AMONG THE BLACK TIMING LINES

Name ______________________________ School ______________________
last first middle

Age ______ Grade ________ ☐ male ☐ female Date ______________________
year day month year

I.D. number ______________________

1. Vocabulary Test
VA-CF.
2. Mathematics Aptitude Test
RA-CF.

Example • Be sure that all points distributed add up to 10.

1	2	3	4	5
6	1	3	0	0

Vocabulary Test -- VA-CF.

1.	1	2	3	4	5	11.	1	2	3	4	5	21.	1	2	3	4	5
2.	1	2	3	4	5	12.	1	2	3	4	5	22.	1	2	3	4	5
3.	1	2	3	4	5	13.	1	2	3	4	5	23.	1	2	3	4	5
4.	1	2	3	4	5	14.	1	2	3	4	5	24.	1	2	3	4	5
5.	1	2	3	4	5	15.	1	2	3	4	5	25.	1	2	3	4	5
6.	1	2	3	4	5	16.	1	2	3	4	5	26.	1	2	3	4	5
7.	1	2	3	4	5	17.	1	2	3	4	5	27.	1	2	3	4	5
8.	1	2	3	4	5	18.	1	2	3	4	5	28.	1	2	3	4	5
9.	1	2	3	4	5	19.	1	2	3	4	5	29.	1	2	3	4	5
10.	1	2	3	4	5	20.	1	2	3	4	5	30.	1	2	3	4	5

Please keep this answer sheet for one more test.

Mathematics Aptitude Test -- RA-CF.

1.	1	2	3	4	5	6.	1	2	3	4	5	11.	1	2	3	4	5
2.	1	2	3	4	5	7.	1	2	3	4	5	12.	1	2	3	4	5
3.	1	2	3	4	5	8.	1	2	3	4	5	13.	1	2	3	4	5
4.	1	2	3	4	5	9.	1	2	3	4	5	14.	1	2	3	4	5
5.	1	2	3	4	5	10.	1	2	3	4	5	15.	1	2	3	4	5

Please place this answer sheet inside the test booklet
and hand them in.

Name ______________________________ School ____________________
last first middle

149

Age ________ Grade ________ ☐ male ☐ female Date ____________________
year day month year

I.D. number ____________________

1. Vocabulary Test
   VA-EL.
2. Mathematics Aptitude Test
   RA-EL.

Example * Cross out choices which you believe are incorrect.

~~1~~ ~~2~~ ~~3~~ ~~4~~ 5

Vocabulary Test -- VA-EL.

1. 1 2 3 4 5
2. 1 2 3 4 5
3. 1 2 3 4 5
4. 1 2 3 4 5
5. 1 2 3 4 5
6. 1 2 3 4 5
7. 1 2 3 4 5
8. 1 2 3 4 5
9. 1 2 3 4 5
10. 1 2 3 4 5
11. 1 2 3 4 5
12. 1 2 3 4 5
13. 1 2 3 4 5
14. 1 2 3 4 5
15. 1 2 3 4 5
16. 1 2 3 4 5
17. 1 2 3 4 5
18. 1 2 3 4 5
19. 1 2 3 4 5
20. 1 2 3 4 5
21. 1 2 3 4 5
22. 1 2 3 4 5
23. 1 2 3 4 5
24. 1 2 3 4 5
25. 1 2 3 4 5
26. 1 2 3 4 5
27. 1 2 3 4 5
28. 1 2 3 4 5
29. 1 2 3 4 5
30. 1 2 3 4 5

Please keep this answer sheet for one more test.

Mathematics Aptitude Test -- RA-EL.

1. 1 2 3 4 5
2. 1 2 3 4 5
3. 1 2 3 4 5
4. 1 2 3 4 5
5. 1 2 3 4 5
6. 1 2 3 4 5
7. 1 2 3 4 5
8. 1 2 3 4 5
9. 1 2 3 4 5
10. 1 2 3 4 5
11. 1 2 3 4 5
12. 1 2 3 4 5
13. 1 2 3 4 5
14. 1 2 3 4 5
15. 1 2 3 4 5

Please place this answer sheet inside the test booklet
and hand them in.

NAME ______________________________
Last First Middle

FACULTY or SCHOOL ______________________________

AGE ______ Years YEAR or GRADE ____________ ☐ Male ☐ Female DATE ____________ Day Month Year

I. D. NUMBER

0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9

150

1. VB TEST

2. RB TEST

Indicate response by placing a mark between the guidelines as shown in the example. Use HB pencil. Don't make marks longer than guidelines.

Example

1 2 3 4 5

Vocabulary Test -- VB.

* This test has only 4 choices.

1. 1 2 3 4
2. 1 2 3 4
3. 1 2 3 4
4. 1 2 3 4
5. 1 2 3 4
6. 1 2 3 4
7. 1 2 3 4
8. 1 2 3 4
9. 1 2 3 4
10. 1 2 3 4
11. 1 2 3 4
12. 1 2 3 4
13. 1 2 3 4
14. 1 2 3 4
15. 1 2 3 4
16. 1 2 3 4
17. 1 2 3 4
18. 1 2 3 4
19. 1 2 3 4
20. 1 2 3 4
21. 1 2 3 4
22. 1 2 3 4
23. 1 2 3 4
24. 1 2 3 4
25. 1 2 3 4
26. 1 2 3 4
27. 1 2 3 4
28. 1 2 3 4
29. 1 2 3 4
30. 1 2 3 4

Please keep this answer sheet for one more test.

Mathematics Aptitude Test -- RB.

1. 1 2 3 4 5
2. 1 2 3 4 5
3. 1 2 3 4 5
4. 1 2 3 4 5
5. 1 2 3 4 5
6. 1 2 3 4 5
7. 1 2 3 4 5
8. 1 2 3 4 5
9. 1 2 3 4 5
10. 1 2 3 4 5
11. 1 2 3 4 5
12. 1 2 3 4 5
13. 1 2 3 4 5
14. 1 2 3 4 5
15. 1 2 3 4 5

Please place this answer sheet inside the test booklet and hand them in.

IBM5047

OVERPRINT ANSWER SHEET

TABLE 72

NUMBER OF UNIVERSITY STUDENTS RATING FOR OPTION WEIGHTS

Class	Number
Ed. Psy. 487	8
Ed. Psy. 489	23
Ed. Psy. 504	13
Total	44

TABLE 73

OPTION WEIGHTS FOR VOCABULARY TEST: FORM A

Item	Option				
	1	2	3	4	5
1	1.977	1.909	3.136	5.000	3.091
2	3.182	2.682	2.364	2.032	5.000
3	2.032	5.000	2.455	2.909	2.864
4	2.909	2.864	2.159	5.000	2.500
5	1.909	2.523	5.000	3.591	2.227
6	1.818	5.000	3.341	2.773	2.364
7	1.705	2.045	5.000	3.750	2.818
8	3.091	2.750	5.000	2.295	2.568
9	5.000	3.227	2.545	2.432	2.318
10	2.682	3.068	5.000	2.523	2.227
11	3.091	2.591	5.000	2.682	2.477
12	1.636	5.000	2.364	3.318	2.818
13	2.477	1.841	2.727	5.000	3.318
14	3.136	2.841	5.000	2.773	1.909
15	2.432	3.159	5.000	2.841	1.682
16	2.659	2.250	2.182	5.000	3.318
17	5.000	3.568	2.886	2.273	1.523
18	5.000	2.955	1.932	2.636	2.886
19	2.682	5.000	2.773	2.364	2.864
20	2.364	2.841	2.636	2.705	5.000
21	2.795	2.886	2.727	2.295	5.000
22	2.659	1.750	2.795	3.227	5.000
23	2.545	2.636	2.727	5.000	2.477
24	5.000	2.818	2.955	2.591	2.500
25	2.409	1.477	5.000	3.045	3.295

TABLE 74

OPTION WEIGHTS FOR MATHEMATICS APTITUDE TEST: FORM A

Item	Option				
	1	2	3	4	5
1	2.273	3.318	2.205	2.750	5.000
2	2.864	2.682	5.000	3.386	2.045
3	3.341	5.000	3.136	2.409	2.114
4	2.636	3.477	5.000	2.568	1.841
5	2.000	3.091	3.136	5.000	2.545
6	5.000	3.409	2.886	2.182	1.886
7	1.841	2.659	5.000	3.364	2.682
8	2.068	2.341	2.795	3.295	5.000
9	2.773	5.000	3.205	2.568	2.205
10	2.409	3.591	5.000	3.273	1.932
11	5.000	3.091	2.864	2.477	2.341
12	1.955	2.614	3.250	5.000	3.409
13	2.023	2.500	2.955	3.182	5.000
14	1.864	2.568	2.841	5.000	3.432
15	2.955	5.000	2.636	2.545	2.294

APPENDIX B

NORMS FOR VOCABULARY TEST: FORM B AND MATHEMATICS APTITUDE TEST: FORM B

TABLE 75

NORMS FOR VOCABULARY TEST: FORM B

Raw Score	Percentile Score	T-Score
30	99.9	81.0
29	99.7	78.1
28	99.5	75.8
27	99.1	73.3
26	98.4	71.5
25	97.3	69.3
24	95.5	67.0
23	93.0	64.8
22	89.5	62.5
21	84.5	60.2
20	78.0	57.7
19	71.0	55.5
18	63.0	53.3
17	54.0	51.0
16	45.0	48.7
15	36.0	46.4
14	28.5	44.3
13	21.0	42.0
12	15.0	39.6
11	10.5	37.5
10	7.0	35.3

TABLE 75 (continued)

Raw Score	Percentile Score	T-Score
9	4.5	33.0
8	2.7	30.7
7	1.6	28.5
6	0.9	26.4
5	0.5	24.3
4	0.25	21.9
3	0.12	19.7
2	0.05	17.0
1	0.025	15.0
0	0.01	13.0

Number of students (grade 9) : 1028
Number of test items : 30
Number of item responses : 4
Mean of the total group : 16.25
Standard deviation : 4.05

Prepared by Wanlop Kansup
Department of Ed. Psy.
The University of Alberta
March, 1973.

TABLE 76

NORMS FOR MATHEMATICS APTITUDE TEST: FORM B

Raw Score	Percentile Score	T-Score
15	99.9	87.0
14	99.9	83.0
13	99.8	78.8
12	99.3	74.5
11	97.7	70.0
10	95.0	66.5
9	87.5	62.0
8	77.0	57.4
7	62.0	53.1
6	46.0	49.0
5	30.0	44.8
4	18.5	41.0
3	9.0	36.6
2	4.0	32.5
1	1.4	28.0
0	0.5	24.3

Number of students (grade 9) : 1028
Number of test items : 15
Number of item responses : 5
Mean of the total group : 5.53
Standard deviation : 2.61

Prepared by Wanlop Kansup
Department of Ed. Psy.
The University of Alberta
March, 1973.

APPENDIX C

INTERCORRELATIONS, MEANS AND STANDARD DEVIATIONS

TABLE 77

INTERCORRELATIONS AMONG SCORES IN CONVENTIONAL TESTING GROUP

	VA_1 -DWS	VA_2 -DWS	RA_1 -DWS	RA_2 -DWS	VA_1 -CVS_1	VA_2 -CVS_1	RA_1 -CVS_1	RA_2 -CVS_1	VB	RB	LA	MA	SC	AV
VA_1 -DWS	-	0.620	0.194	0.237	0.965	0.662	0.345	0.359	0.584	0.312	0.480	0.442	0.464	0.517
VA_2 -DWS		-	0.203	0.262	0.620	0.943	0.331	0.386	0.639	0.360	0.438	0.387	0.449	0.491
RA_1 -DWS			-	0.530	0.167	0.182	0.807	0.448	0.133	0.390	0.124	0.259	0.192	0.234
RA_2 -DWS				-	0.213	0.256	0.467	0.808	0.244	0.481	0.145	0.300	0.257	0.268
VA_1 -CVS_1					-	0.680	0.333	0.346	0.589	0.302	0.486	0.445	0.472	0.517
VA_2 -CVS_1						-	0.333	0.404	0.670	0.376	0.471	0.410	0.484	0.518
RA_1 -CVS_1							-	0.607	0.277	0.495	0.271	0.470	0.395	0.422
RA_2 -CVS_1								-	0.383	0.562	0.296	0.514	0.461	0.475
VB									-	0.393	0.470	0.438	0.442	0.505
RB										-	0.235	0.382	0.348	0.368
LA											-	0.585	0.604	0.799
MA												-	0.710	0.825
SC													-	0.838
AV														-

TABLE 78

INTERCORRELATIONS AMONG SCORES IN CONFIDENCE TESTING GROUP

	VA 1					VA 2					RA 1					RA 2					VB	RB	LA	MA	SC	AV
	CWS 1	CWS 2	CWS 3	CWS 4	CWS 5	CWS 1	CWS 2	CWS 3	CWS 4	CWS 5	CWS 1	CWS 2	CWS 3	CWS 4	CWS 5	CWS 1	CWS 2	CWS 3	CWS 4	CWS 5						
VA$_1$-CWS$_1$	-	983	985	972	996	764	727	771	768	756	455	449	461	461	448	407	390	420	424	402	646	329	336	333	400	410
-CWS$_2$		-	939	916	986	735	715	727	718	731	452	452	453	450	448	408	395	416	418	403	638	338	342	332	396	408
-CWS$_3$			-	997	974	770	718	792	794	757	445	433	457	460	436	397	375	415	421	390	636	307	323	332	397	406
-CWS$_4$				-	957	765	708	792	797	750	434	420	449	453	423	386	362	407	414	379	627	295	312	320	386	393
-CWS$_5$					-	756	723	759	754	753	454	450	457	455	450	403	388	412	414	400	648	328	334	332	392	407
VA$_2$-CWS$_1$						-	978	980	968	991	459	451	463	462	454	448	434	455	457	444	673	335	426	392	468	486
-CWS$_2$							-	933	910	983	450	449	447	443	448	441	437	440	439	441	672	318	450	390	473	497
-CWS$_3$								-	997	972	444	433	454	456	438	437	417	452	458	430	647	314	388	383	453	462
-CWS$_4$									-	955	440	426	452	455	431	432	409	451	458	423	630	313	374	377	445	450
-CWS$_5$										-	455	451	456	452	454	443	435	445	445	443	672	312	419	385	456	479
RA$_1$-CWS$_1$											-	992	992	984	997	669	648	679	680	665	429	488	304	499	426	466
-CWS$_2$												-	971	957	994	661	648	665	663	660	429	475	304	489	417	458
-CWS$_3$													-	998	985	666	640	684	687	660	426	493	300	502	433	469
-CWS$_4$														-	973	662	632	683	688	654	422	494	297	499	433	467
-CWS$_5$															-	668	650	675	674	667	427	483	303	499	421	464
RA$_1$-CWS$_1$																-	992	993	986	997	393	530	274	479	431	441
-CWS$_2$																	-	971	959	994	390	512	272	469	422	435
-CWS$_3$																		-	999	985	390	537	272	483	435	442
-CWS$_4$																			-	975	386	539	271	479	434	439
-CWS$_5$																				-	394	522	275	484	429	442
VB																					-	345	394	352	447	476
RB																						-	192	319	331	306
LA																							-	595	608	795
MA																								-	707	835
SC																									-	851
AV																										-

Note All values are decimal.

TABLE 79

INTERCORRELATIONS AMONG SCORES IN ELIMINATION TESTING GROUP

	VA$_1$ -ELS	VA$_2$ -ELS	RA$_1$ -ELS	RA$_2$ -ELS	VA$_1$ -CVS$_2$	VA$_2$ -CVS$_2$	RA$_1$ -CVS$_2$	RA$_2$ -CVS$_2$	VB	RB	LA	MA	SC	AV
VA$_1$ -ELS	-	0.671	0.359	0.310	0.769	0.602	0.317	0.305	0.672	0.285	0.395	0.350	0.432	0.435
VA$_2$ -ELS		-	0.341	0.327	0.556	0.827	0.322	0.312	0.664	0.272	0.382	0.326	0.431	0.430
RA$_1$ -ELS			-	0.599	0.216	0.227	0.938	0.573	0.384	0.383	0.355	0.556	0.405	0.462
RA$_2$ -ELS				-	0.191	0.208	0.577	0.936	0.381	0.360	0.344	0.499	0.395	0.452
VA$_1$ -CVS$_2$					-	0.663	0.245	0.237	0.546	0.236	0.254	0.214	0.324	0.285
VA$_2$ -CVS$_2$						-	0.254	0.259	0.583	0.224	0.245	0.175	0.273	0.249
RA$_1$ -CVS$_2$							-	0.589	0.361	0.387	0.325	0.533	0.385	0.432
RA$_2$ -CVS$_2$								-	0.354	0.369	0.292	0.463	0.351	0.394
VB									-	0.307	0.471	0.347	0.421	0.455
RB										-	0.203	0.316	0.253	0.242
LA											-	0.622	0.611	0.817
MA												-	0.680	0.811
SC													-	0.820
AV														-

TABLE 80

MEANS AND STANDARD DEVIATIONS OF TEST SCORES UNDER CONVENTIONAL TESTING AND SCORING METHODS

Test	Mean	Standard Deviation
VA_1	12.0	3.5
VA_2	12.6	3.7
RA_1	6.7	2.7
RA_2	6.8	2.8
VB	16.0	5.1
RB	5.6	2.7